Middlesex University London

Middlesex University Research Repository

An open access repository of

Middlesex University research

http://eprints.mdx.ac.uk

Johnstone, Kevin Kennedy (1991) On the nature and effect of power distribution noise in CMOS digital integrated circuits. PhD thesis, Middlesex Polytechnic. [Thesis]

Final accepted version (with author's formatting)

This version is available at: https://eprints.mdx.ac.uk/13368/

Copyright:

Middlesex University Research Repository makes the University's research available electronically.

Copyright and moral rights to this work are retained by the author and/or other copyright owners unless otherwise stated. The work is supplied on the understanding that any use for commercial gain is strictly forbidden. A copy may be downloaded for personal, non-commercial, research or study without prior permission and without charge.

Works, including theses and research projects, may not be reproduced in any format or medium, or extensive quotations taken from them, or their content changed in any way, without first obtaining permission in writing from the copyright holder(s). They may not be sold or exploited commercially in any format or medium without the prior written permission of the copyright holder(s).

Full bibliographic details must be given when referring to, or quoting from full items including the author's name, the title of the work, publication details where relevant (place, publisher, date), pagination, and for theses or dissertations the awarding institution, the degree type awarded, and the date of the award.

If you believe that any material held in the repository infringes copyright law, please contact the Repository Team at Middlesex University via the following email address:

eprints@mdx.ac.uk

The item will be removed from the repository while any claim is being investigated.

See also repository copyright: re-use policy: http://eprints.mdx.ac.uk/policies.html#copy

Middlesex University London

Middlesex University Research Repository:

an open access repository of
Middlesex University research

http://eprints.mdx.ac.uk

Johnstone, Kevin Kennedy, 1991.
On the nature and effect of power distribution noise in CMOS digital integrated circuits.
Available from Middlesex University's Research Repository.

Copyright:

Middlesex University Research Repository makes the University's research available electronically.

Copyright and moral rights to this thesis/research project are retained by the author and/or other copyright owners. The work is supplied on the understanding that any use for commercial gain is strictly forbidden. A copy may be downloaded for personal, non-commercial, research or study without prior permission and without charge. Any use of the thesis/research project for private study or research must be properly acknowledged with reference to the work's full bibliographic details.

This thesis/research project may not be reproduced in any format or medium, or extensive quotations taken from it, or its content changed in any way, without first obtaining permission in writing from the copyright holder(s).

If you believe that any material held in the repository infringes copyright law, please contact the Repository Team at Middlesex University via the following email address:
eprints@mdx.ac.uk

The item will be removed from the repository while any claim is being investigated.

ON THE NATURE AND EFFECT OF POWER DISTRIBUTION NOISE IN CMOS DIGITAL INTEGRATED CIRCUITS

A thesis submitted to the Council for National Academic Awards

by

Kevin Kennedy Johnstone B.Sc.(Edin), M.Sc.(Edin)

in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Y942829X

SITE HE	MIDDLESEX POLYTECHNIC LIBRARY
ACCESSION No.	7185038-4
CLASS No.	621.38224 JOH
SPECIAL COLLECTION	Thesis Collection

November 1991

Middlesex Polytechnic

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 3

Chapter 4

Chapter 4

Chapter 5

ON THE NATURE AND EFFECT OF POWER DISTRIBUTION NOISE IN CMOS DIGITAL INTEGRATED CIRCUITS

Kevin Kennedy Johnstone

- ABSTRACT-

The thesis reports on the development of a novel simulation method aimed at modelling power distribution noise generated in digital CMOS integrated circuits.

The simulation method has resulted in new information concerning:

1. The magnitude and nature of the power distribution noise and its dependence on the performance and electrical characteristics of the packaged integrated circuit. Emphasis is laid on the effects of resistive, capacitative and inductive elements associated with the packaged circuit.

2. Power distribution noise associated with a generic systolic array circuit comprising 1,020,000 transistors, of which 510,000 are synchronously active. The circuit is configured as a linear array which, if fabricated using two-micron bulk CMOS technology, would be over eight centimetres long and three millimetres wide. In principle, the array will perform 1.5 X 10^{11} operations per second.

3. Power distribution noise associated with a non-array-based signal processor which, if fabricated in 2-micron bulk CMOS technology, would occupy 6.7 sq.cm. The circuit contains about 900,000 transistors, of which 600,000 are functional and about 300,000 are used for yield enhancement. The processor uses the RADIX-2 algorithm and is designed to achieve 2 X 10^{8} floating point operations per second.

4. The extent to which power distribution noise limits the level of integration and/or performance of such circuits using standard and non-standard fabrication and packaging technology.

5. The extent to which the predicted power distribution noise levels affect circuit susceptibility to transient latch-up and electromigration.

It concludes the nature of CMOS digital integrated circuit power distribution noise and recommends ways in which it may be minimised.

It outlines an approach aimed at mechanising the developed simulation methodology so that the performance of power distribution networks may more routinely be assessed.

Finally, it questions the long term suitability of mainly digital techniques for signal processing.

ACKNOWLEDGEMENTS

I thank Professor John Butcher of the Microelectronics Centre at Middlesex Polytechnic for affording me the research fellowship and for his able guidance throughout my research programme.

I thank Andrew Stewart, formerly of Plessey Research (Caswell) Limited, for a concise description of his signal processing architecture and for his helpful assistance in general.

I thank Dr. Will Moore of the Department of Engineering Science at the University of Oxford for his advice and support at various stages of the programme.

I thank Dick Pearson of Texas Instruments Limited for his expert preparation of all illustrations and text.

This research programme was funded by the UK Alvey Directorate together with Middlesex Polytechnic.

INTRODUCTION

1. Research Objectives

The objectives of this research programme are as follows.

(i) *To develop an accurate and flexible method of assessing the degree to which power distribution noise is generated in CMOS digital integrated circuits.*

(ii) *To determine the magnitude of the power distribution noise and its dependence on the performance and electrical characteristics of the packaged integrated circuit.*

(iii) *To assess the extent to which power distribution noise limits the achievable level of synchronous circuit activity for CMOS digital integrated circuits.*

2. Scope of Thesis

The programme is concerned with an array-based and a non-array-based digital signal processor. More specifically, it is concerned with a generic bit-level systolic array processor and a non-array-based 32-bit floating-point processor conforming to IEEE standard 754. For reasons given below, it is assumed that each is fabricated with two-micron bulk CMOS technology.

These architectures were chosen for the following reasons.

The systolic array has a simple structure which is extremely regular and is often cited as a good architecture [a01], [a02] with which to develop yield-enhanced circuits. In addition, its simple, regular structure is amenable to the development of a noise modelling technique.

The chosen non-array-based architecture is a prototype for a commercially available yield-enhanced integrated circuit. At the time of defining the research programme, this prototype was the only such fault-tolerant circuit in Europe. It represented the highest level of monolithic integration for a digital signal processing circuit.

The analysis is concerned with power distribution noise associated with core circuitry and not with peripheral I/O drivers.

The analysis will not be concerned with the effects of resistive, capacitative or inductive elements which are outside the integrated circuit package. An excellent treatment of this problem has been undertaken recently by Keenan [a03].

This analysis will focus on the effects of resistive, capacitative and inductive elements associated with the circuit and with the package itself. In short, it will undertake an

analysis of the primary source of power distribution noise associated with digital computing systems.

Fabrication-related transistor and metallisation parameters were derived from manufacturing data. Accurate data relating only to two-micron bulk CMOS were available.

3. Thesis Structure

Chapter one introduces the technological issues which limit the electrical performance of packaged integrated circuits. The issues addressed are: 1) the electrical characteristics of interconnect and interconnect modelling; 2) clock distribution 3) the electrical characteristics of integrated circuit packages; and 4) power distribution noise.

The objective of chapter two is to review all published analyses of power distribution noise associated with CMOS digital integrated circuits.

Of these analyses [c01], [c02] and [c03] are concerned with simultaneous output driver switching, leaving only [c04], [c05] and [c06] which undertake detailed analyses of power distribution noise associated with core sections of CMOS integrated circuits.

The objective of chapter three is to introduce array-based architectures, systolic architectures in particular, as used to address compute-bound problem solving in digital signal processing and to describe the development of a simulation model used to predict power distribution noise for such highly synchronous architectures.

Chapter three then describes the nature of the power distribution noise and how the model predictions were used to assess the extent to which power distribution technology limits the performance and levels of integration of such circuits. The model is able to predict the power distribution noise associated with a generic systolic array circuit comprising 1,020,000 devices, of which 510,000 are synchronously active. The circuit is configured as a linear array which, if fabricated using two-micron bulk CMOS technology,would be over eight centimetres long and three millimetres wide.

Chapter four is an account of how the simulation methodology, developed in chapter three, was applied to the non-array-based signal processor referred to above. In contrast to the analysis of chapter three, it is not justifiable in the case of a non-array-based architecture to assume that any single circuit dominates the power distribution noise, thus necessitating a detailed analysis of each circuit sub-block and the conditions applicable to each operational mode. In two-micron bulk CMOS technology, the circuit would occupy about 6.7 sq.cm. It contains about 900,000 transistors of which about 600,000 are functional and 300,000 are used for yield enhancement.

Chapter five has two sections. The first is concerned with latch-up and the second with electromigration. This chapter presents an analysis of the extent to which the power distribution noise levels, predicted in chapters three and four, affect circuit

susceptibility to transient latch-up and electromigration.

4. Statement of Originality

The programme of work outlined in this thesis was inspired by Itoh, Nakagawa, Sakui, Horiguchi and Ogura [c05] and by Lea [a04].

As far as can be ascertained, it has resulted in: 1) a novel simulation method which has proven useful in the assessment of power distribution noise; and 2) new information concerning the nature of power distribution noise and the extent to which it may limit the achievable level of synchronous circuit activity for CMOS digital integrated circuits.

1 TECHNOLOGICAL CONSTRAINTS

1.1 Introduction

The objective of this chapter is to introduce the technological issues which have constrained, and those that may constrain, the level of synchronous activity associated with CMOS digital integrated circuits.

The issues addressed in this chapter are the electrical characteristics of interconnect, interconnect modelling, clock distribution, the electrical characteristics of integrated circuit packages and power distribution noise.

The integrated circuit, invented by Texas Instruments in 1959 [b01], [b02], has undergone continuous development aimed at increasing integration levels. This has been realised through regular reductions in transistor dimensions, increases in the circuit area and by improvements in circuit packaging [b03], [b04]. Between 1959 and 1983 transistor dimensions were reduced by eleven percent per year on average and circuit area was increased by nineteen percent per year on average [b05]. These two factors combine to yield an increase in the level of integration in excess of fifty-fold per decade.

Though desirable, this high growth rate has not been sustained in recent years and may fast be approaching a technologically-constrained limit. What are the technological issues which determine the maximum level of circuit integration? The question may be put more precisely: what are the issues which determine the maximum level of *synchronous circuit activity*?

Throughout the development of the integrated circuit, the main factor that has determined the achievable level of circuit integration is fabrication integrity. Reductions in transistor dimensions and increases in circuit area each have increased circuit susceptibility to process and packaging imperfections leading to a lower proportion of functional devices or lower manufacturing yield.

The maximum level of synchronous circuit activity clearly is dependent on transistor size, circuit size and packaging technology. Until recently, the sole metric which has influenced their choice is manufacturing yield. As manufacturing yield improves, allowing the manufacture of circuits with many more devices, circuit-level electrical performance may emerge as an additional factor which must be analysed if the full potential of the integrated circuit is to be realised.

1.2 Interconnect Modelling

1.2.1 Interconnect scaling

Interconnect elements no longer may be considered to have negligible impact on the performance of integrated circuits. With reductions in circuit feature size and

increases in circuit dimensions, parasitic interconnect capacitance and resistance are increased to effect delays which are comparable with gate delays.

Gate delays are decreased because transistor gate, gate-drain overlap and source-drain areas are reduced. Interconnect capacitance is increased as a consequence of an increase in circuit size. This typically more than compensates the reduction in interconnect capacitance per unit length as a result of reduced interconnect track widths.

Interconnect resistance is increased as a result of an increase in circuit size and as a result of reduced cross-sectional area. Interconnect resistivity is independent of cross-sectional area unless this is reduced to a level which is similar to the mean-free-path of the electron. For aluminium, mean-free-paths are less than 100nm and consequently it is reasonable to assume no dependence.

Interconnect capacitance can be modelled by an area term and a periphery term. As shown in Figure 1.1, interconnect capacitance can be modelled by a combination of a parallel plate capacitor of width W_{int} and a cylindrical wire of diameter H_{int} [b06]. In Figure 1.2, C_{int} is shown plotted as a function of W_{int}/t_{field} for two H_{int}/t_{field} ratios where t_{field} is the transistor field oxide thickness. It can be seen that C_{int} approaches an asymptote of 1pF/cm when $W_{int} \approx t_{field} \approx H_{int}$. This is due to peripheral or "fringe" effects.

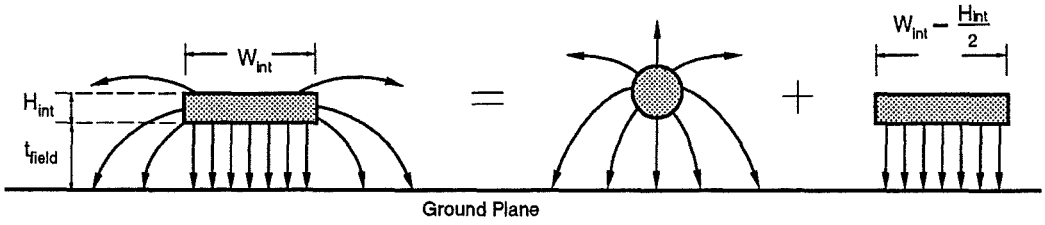


Fig 1.1 Interconnect Capacitance - Area & Periphery Terms

A third capacitance term is that associated with coupling to neighbouring interconnect lines. This coupling term has the undesirable effects of degrading switching speed and enhancing circuit noise thereby degrading significantly circuit performance.

It is clear, from Figure 1.2, that as circuit features are made smaller and circuit sizes are increased, area effects increasingly are dominated by periphery and coupling effects. The problem of analysis therefore grows from one of capacitance per unit length to capacitance per unit area and capacitance per unit volume.

Multi-dimensional modelling aimed at accurately predicting interconnect capacitance has been reported extensively [b07], [b08], [b09].

1.2.2 Distributed RC Lines

The response of a *lumped* RC network to a step potential is

$$V(t) = 1 - \exp(-t/RC)$$

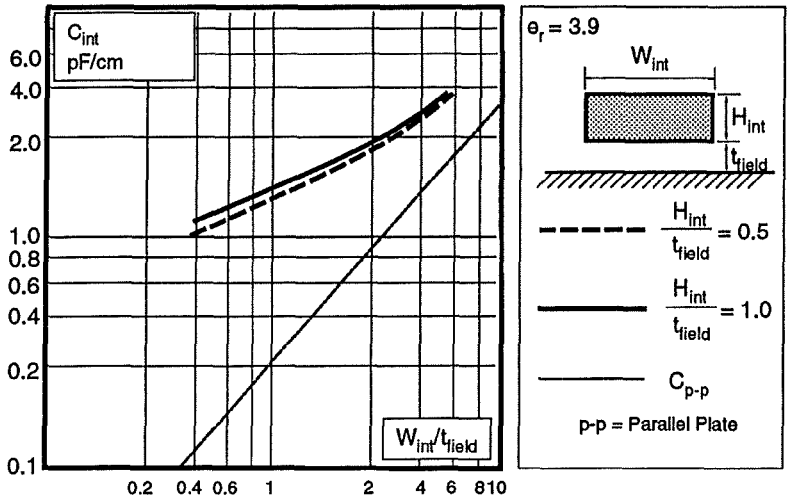


Fig 1.2 Interconnect Capacitance vs Interconnect Width/Oxide Thickness

The time-domain solution of a "distributed" RC network does not exist in closed form. Solutions valid for t<<RC and t>>RC can be obtained by using an approximate expansion for the hyperbolic cosine function associated with the frequency-domain solution.

$$\cosh(x) = \exp(x)/2 \qquad \text{for } x>>1$$

and

$$\cosh(x) = 1+x^2/2! + x^3/4! \qquad \text{for } x<<1$$

The approximation for large x thus is appropriate for high frequency signal components; for small x, it is appropriate for low frequency signal components. Signal transitions are high frequency components while steady-state values are low frequency.

Such time-domain and frequency-domain solutions are difficult to use, however, and it is known that a distributed RC line can be approximated by a lumped resistor and capacitor network.

This approximation lends itself well to computer-aided circuit simulation and the commonly adopted resistor/capacitor configurations are shown in Figure 1.3. The question most often raised in connection with lumped element RC approximation is: "how many sections are needed accurately to model the performance of the distributed line ?".

Antinone and Brown [b10] set about answering this question by undertaking frequency-domain simulations for a T-network and a 2-, 5- and 10-section ladder network and comparing them with exact mathematical analyses. Their results are shown in Figure 1.4. The authors conclude that the number of lumped elements should be chosen so that the time constant of each ($r_n.c_n$) is one-tenth the time constant associated with the interconnect section being studied.

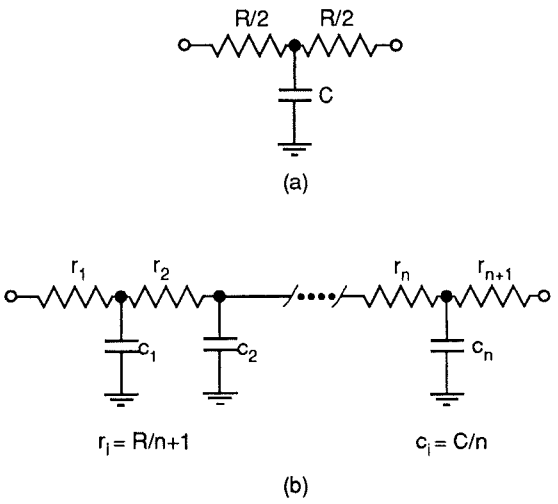


Fig 1.3 T-network & Ladder Network distributed RC models

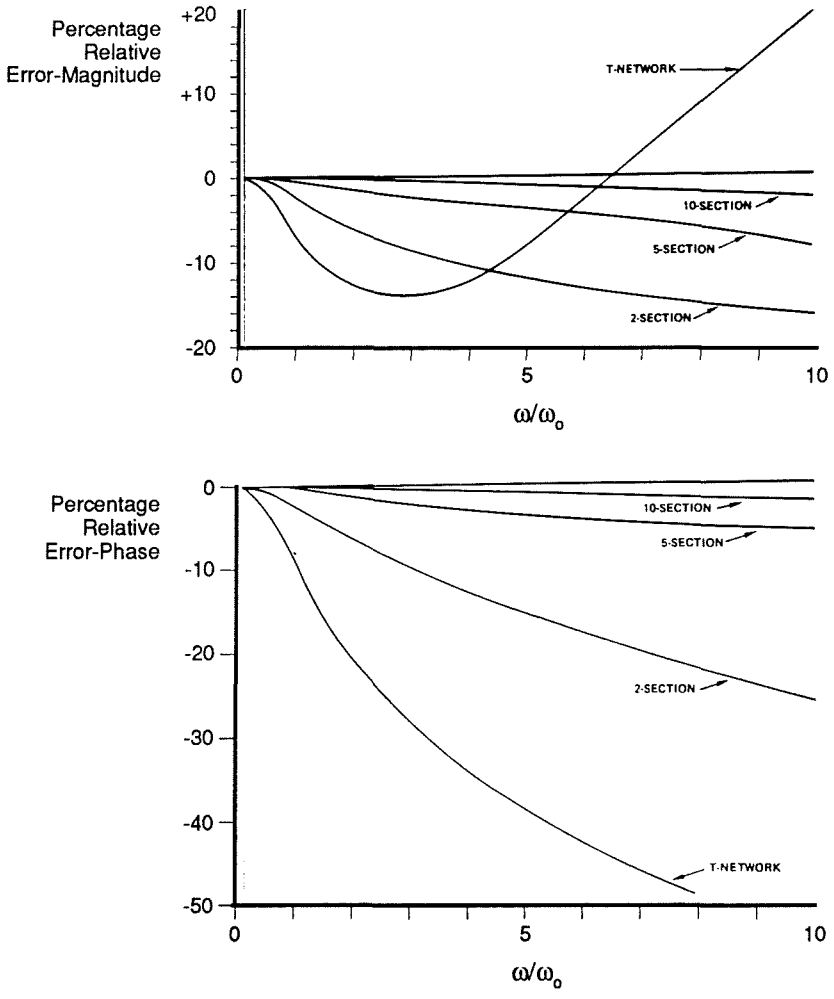


Fig 1.4 Error of Simulated Results Relative to Theoretical as a Function of Frequency

1.2.3 Transmission Lines

Longer interconnect lines and higher clock frequencies associated with larger integrated circuits constituted by smaller, and consequently faster, devices ultimately will result in clock transition times which are comparable with interconnect delay. At this stage, inductance becomes a significant electrical characteristic of integrated circuit interconnections and therefore introduces the need for transmission line modelling.

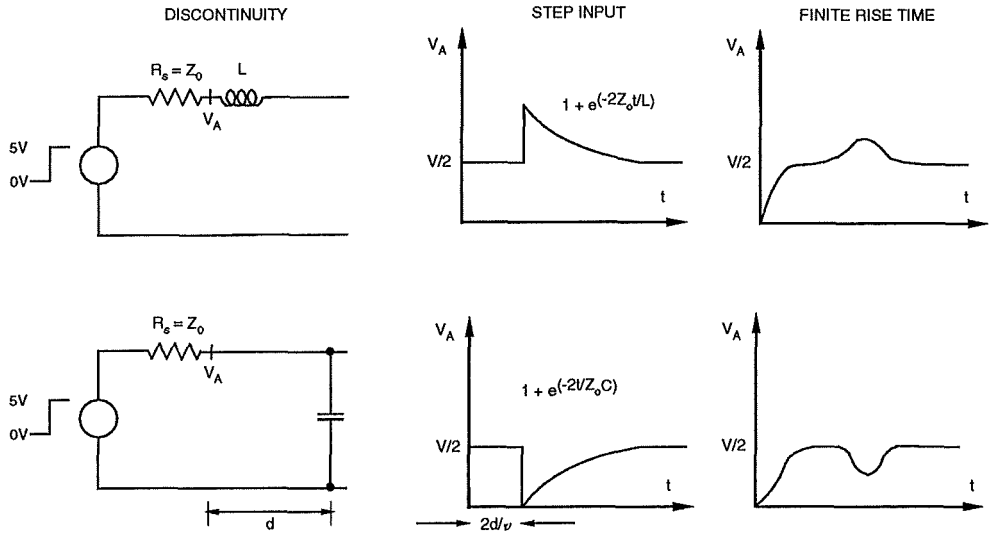


Fig 1.5 Inductive & Capacitive Discontinuities

To obtain accurate inter-circuit delay and noise estimates, bond wires, package parasitics and circuit board interconnections are modelled as transmission lines. At the circuit board level, signal reflections can be generated at capacitative and inductive discontinuities due to interconnect lines, connectors, package pins, vias and corners associated with board wiring.

As an example of such phenomena, Figure 1.5 illustrates signal reflections associated with a capacitative and an inductive discontinuity. The transmission line is driven by a signal generator with source resistance $R_S=Z_0$ and the reflected signal is observed at the source. It is assumed that the line is infinitely long so that all reflections are due to the discontinuity.

Immediately after the signal reaches the discontinuity, the capacitor has low impedance and shorts the line. Later its impedance is restored and the voltage disturbance decays with a time constant $Z_0C/2$.

The inductor, on the other hand, has high impedance immediately after the signal reaches the discontinuity and the voltage step doubles. Later, as current increases, its impedance falls and the voltage disturbance decays with a time constant of $L/2Z_0$.

In the second set of signals, shown in Figure 1.5, it is assumed that the input step has a rise time which is larger than the discontinuity time constants. Then there is insufficient time for the disturbances to reach their theoretical maxima.

Rise Time t_r(ps)	Critical Line Length lcrit=vt_r/2.5 (cm)
50	0.3
100	0.6
250	1.5
500	3.0
750	4.5
1000	6.0

Fig 1.6 Critical Transmission Line Lengths

At present, CMOS integrated circuits do not exhibit transmission line effects; they are insufficiently large or fast. How large and fast need they be ? Transmission line behaviour becomes significant when the rise time, t_r, of a signal becomes comparable with, or less than, interconnect delay, t_{int}.

It is assumed that transmission line phenomena become "significant" when interconnect delay is comparable with, or less than, one quarter period of the dominant Fourier component associated with the clock signal. When this is true, it can be shown that $tr < 2.5t_{int}$. Interconnect lengths corresponding to $tr = 2.5t_{int}$, so-called critical transmission line lengths, are listed for CMOS technology in Figure 1.6.

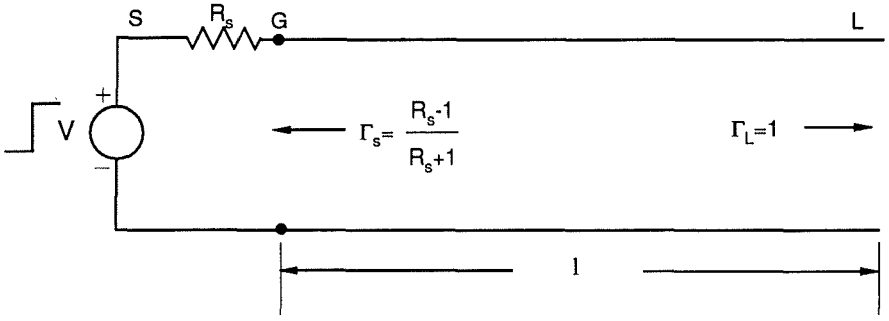


Fig 1.7 Unterminated Transmission Line driven by a Source Resistance R_s

The relationship is complicated further by a dependence on the ratio of signal source resistance Rs to the characteristic impedance of the line Z_0. If the transmission line of Figure 1.7 is modelled as a lumped capacitor, the rise time of the resulting RC network can be expressed as

$$t_r = 2.3R_sC = 2.3R_scl$$

where 'c' is the capacitance per unit length and 'l' is the transmission line length. The delay of the line is expressed as $t_{int} = l/v$ where 'v' is the signal velocity.

Transmission line phenomena would be negligible if the round-trip delay of the line

were much less than the rise time ($2t_{int} << t_r$). For this to be true

$$1/v << R_s cl \Rightarrow 1/v.c << R_s$$

It is thus concluded that

$$Z_o = 1/vc \Rightarrow Zo << R_s$$

An additional requirement on the relationship for transmission line phenomena is that the signal source resistance should be much less than the characteristic impedance of the line.

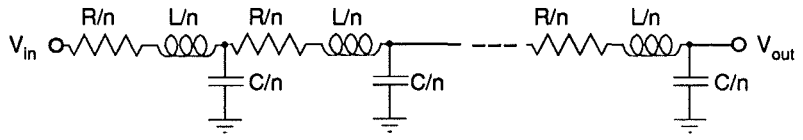


Fig 1.8 Lossy Transmission Line Model

The above analysis applies to loss-less transmission lines. Integrated circuit interconnects, usually of aluminium, have significant resistance and should be modelled as lossy transmission lines. Resistance introduces signal attenuation and it can be shown that for a lossy transmission line of length 'l', the voltage transfer function can be expressed as

$$T = \exp(-R/2Z_o)$$

where R is the total line resistance and Z_o its characteristic impedance.

Figure 1.8 illustrates an approximate model for a lossy transmission line. This can be used easily in circuit simulators. For this case,

$$Z_o = \sqrt{(L/C)};\ \text{and,}\ t_{int} = \sqrt{(LC)}$$

where R, C and L are the total resistance, capacitance and inductance of the line.

Transmission line phenomena for integrated circuit interconnect are complicated further by the "slow wave" effect [b11], [b12]. The explanation of this effect lies in the fact that the semiconductor behaves as a conductor for capacitative effects and as an insulator for inductive effects thereby "distorting" magnetic and electric fields associated with the interconnect.

Integrated circuit transmission line effects have been analysed and reported on extensively [b13], [b14], [b15]. These analyses have been embraced in the development of simulators such as MCLINES, which can be used to predict accurately capacitance and inductance values for the equivalent circuit of Figure 1.8.

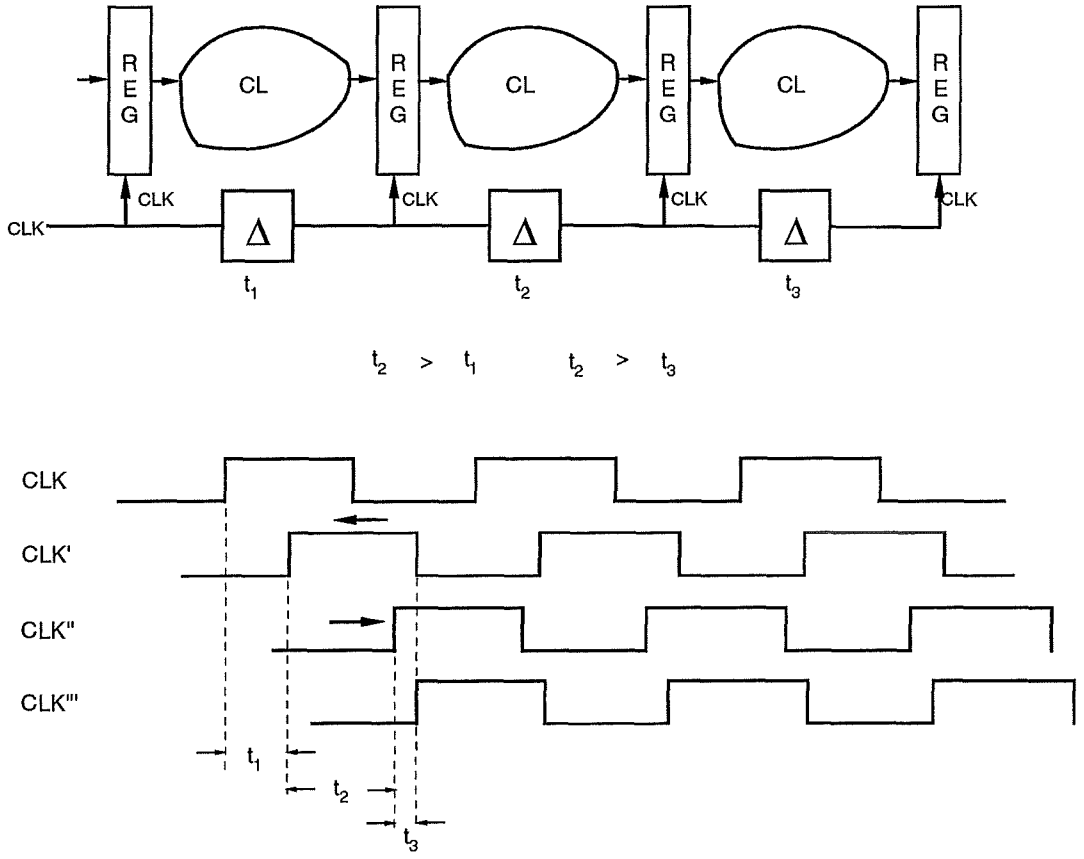


Fig 1.9 Clock Distribution Tuning

Clock distribution is, by implication, of major concern in the design of highly synchronous integrated circuits. It is clear that, if the improved performance associated with smaller devices is to be exploited, then clock periods will be shortened. Similarly, it is clear that as circuit sizes increase, then it will be necessary to distribute the clock further from its source along lossy transmission lines [b16], [b17].

1.3.1 Clock Skew

Clock skew manifests itself in two forms. The first is if two circuits, governed by the same clock, are not equidistant from the clock driver. They consequently have unequal interconnect loading and receive the clock at different times. The second is when process variations have resulted in unequal gate and interconnect delays. For each, the effect is the same: to reduce the degree of synchronous activity in the circuit.

Clearly, as clock periods are reduced, circuits will exhibit increased susceptibility to clock skew. For this reason, it is useful to think of clock skew, expressed as a fraction of the clock period, as a measure of the effectiveness of the clock distribution scheme.

Clock skew is not often controllable by design. If it is, then it can be used to effect an improvement in circuit performance. This is referred to as "tuning" the clock signal or the clock distribution network. For example, if the delays associated with pipeline stages are unequal, circuit performance may be optimised either by retarding the data capture clock for slower pipeline stages or by advancing it for faster stages. This technique is illustrated in Figure 1.9.

Though useful, it must be stressed that the intentional introduction of clock skew necessitates precise skew control, a task more exacting and generally less robust than skew minimisation through advanced distribution schemes. Two common examples of such schemes are locally synchronous and H-tree clock distribution.

1.3.2 Locally Synchronous Distribution

In this scheme the circuit, or system, is partitioned into synchronous islands each with its own independent clock. This technique avoids the distribution of a single global clock which inevitably will lead to unacceptable clock skew and compromised performance.

Once the circuit, or system, is divided into locally synchronous regions, clock skew problems effectively are transferred to the inter-region communication mechanism [b18], [b19], [b20]. There exist two common communication mechanisms: 1) asynchronously using a self-timed discipline; and 2) synchronously using a global clock of lower frequency than the local clocks.

A self-timed discipline is one which is made up of speed-independent circuits which communicate via a set of asynchronous protocols [b18], [b19]. The system changes state only when all circuits signal completion, thereby eliminating the need for a global clock and thus, by definition, the clock skew problem disappears.

Another advantage of self-timed systems is that, unlike synchronous systems, performance is not determined by the worst-case signal delay. In a self-timed system, a computational step is initiated after its sequential predecessors have completed their computational steps. The total delay is the aggregate delay of the computational steps and as such it reflects the average delay instead of the worst-case delay [b20].

If the worst-case path is activated only rarely, a self-timed system will perform better than a synchronous one but at the cost of the extra logic required for communication.

In the case of the second mechanism, the locally synchronous regions use a relatively high frequency clock and the global communication is achieved with a much lower frequency bus. In practice, the bus width is determined by a combination of the system requirements and the maximum supportable global clock frequency.

A possible complication of this communication mechanism is that of metastability introduced by clock skew between local and global clocks [b21], [b22]. This can be avoided through the use of phase-locked loop circuits to adjust dynamically the relative phase of the clocks [b23].

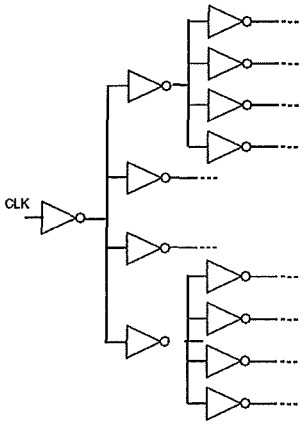


Fig 1.10
Multiple Driver Distribution Scheme

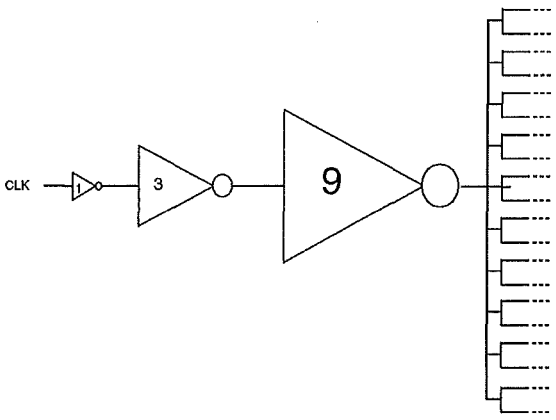


Fig 1.11 Successively Higher Gain Driver Distribution Scheme

The mechanism is flexible in the sense that it may be extended to include a hierarchy of clock frequencies appropriate for large circuits containing small regions of relatively high performance circuitry. This technique has been adopted by Lea and Coleman [b24] in the development of their associative processing architecture.

1.3.3 The H-tree Distribution

Figure 1.10 shows a regular clock distribution scheme used to reduce clock skew due to interconnect. In this scheme, the idea is to use identical clock drivers with equal loads. The principal cause of clock skew is then that of unequal driver delay due to process-related parametric spread. If interconnect delay is relatively insignificant in comparison with delay mismatch, skew can be further reduced by replacing the multiple buffer structure of Figure 1.10 with that of Figure 1.11 in which the tree branches are driven by a cascade of a successively higher gain drivers.

In order that clock skew be minimised, the clock should be distributed in such a way that the clock is electrically equidistant from all functional blocks it serves. Clock signals will be delayed equally on arrival at the interface to the functional block as in the case of the locally synchronous system. Intra-block clock skew will be dependent on block size which therefore can be chosen to be as large as the application can sustain.

The clock distribution scheme shown in the Figure 1.12 [b25] adopts the above ideas by repeating recursively an H-shaped structure. Clearly, this H-tree scheme can be easily extended to a distribution system in which phase-locked loop circuits have been included to minimise process-related variations in regional clock driver delays. This distribution scheme offers minimal clock skew and may be applied to systems ranging from circuit boards through to sub-circuits of an integrated circuit.

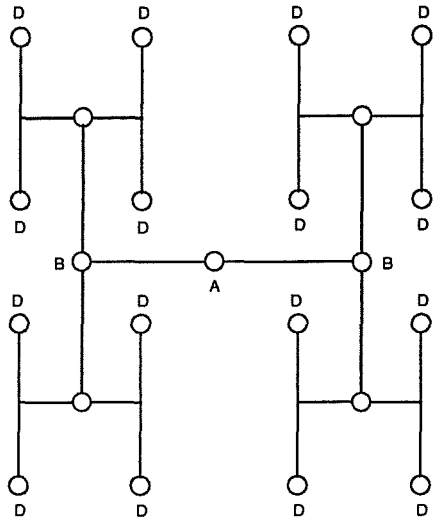


Fig 1.12 H-tree Clock Distribution Scheme

1.4 Packaging Technology

Packaging is required to supply integrated circuits with an electrical interface to the outside world, to remove heat generated by the circuit and to provide the circuit with physical support and environmental protection.

Currently, as much as 50% of computer system delay is attributable to packaging.. This figure has been estimated to increase to 80% by the end of the decade [b26].

In order that system performance be optimised, it will be necessary to ensure: 1) that packages contain larger circuits; and 2), that they are grouped more densely on circuit boards.

It is for these reasons that increased importance must be placed on the electrical and thermal characteristics of integrated circuit packages.

1.4.1 The Dual-in-Line Package

The dual-in-line package derives its name from its pin configuration and remains the most common form of integrated circuit packaging. It is usual to find through-hole board mounting technology used in combination with this form of packaging.

Dual-in-line packaging is formed by bonding the integrated circuit to a bond lead frame. Wire bonds are made subsequently to provide electrical connections from the circuit to the bond leads and the whole device is encapsulated either in ceramic or, more commonly, in plastic.

Dual-in-line packages are inserted through the holes of a circuit board and, to effect permanent electrical contact, the entire bottom surface of the board is dipped in a solder which adheres only to the metal pins, the copper-plated walls of the holes and

the contact areas surrounding the holes.

Through-hole mounting provides robust mechanical support to the circuit and offers high resistance to thermal stress. In combination with dual-in-line packaging, however, it does not provide good packing density. Inherently, the pin configuration of dual-in-line packages gives rise to "pin-bound packages" and through-hole mounting makes inefficient use of board space since it occupies both sides of the circuit board.

The above factors result in relatively lengthy board-level interconnections which, combined with dual-in-line package parasitics, are likely to degrade system performance [b27], [b28].

1.4.2 The Pin Grid Array

An alternative to the dual-in-line package is the pin grid array. Unlike the dual-in-line package, the pin grid array features bond leads on its entire bottom surface. This configuration offers obvious benefits over the dual-in-line package and pin-outs of over 300 are practical with pin grid arrays.

Pin grid arrays are available in two forms: cavity-up and cavity-down depending on whether the pins and circuit are fixed on the same or opposite sides of the, usually square, package. The advantage of the cavity-up form is that a heat sink may be attached to the reverse side of the package while the advantage of cavity-down is that one entire side may be occupied by pins.

To support a high pin-out, pin grid arrays require multilayer ceramic, usually alumina, substrates. This requirement results in pin grid arrays having two undesirable features. They have high bond lead parasitics because of the high dielectric constant of alumina, and they are subject to mechanical stress at the circuit board interface. On the positive side, ceramics do exhibit high thermal conductivity and provide good hermiticity.

As an alternative to multilayer ceramic substrates, pin grid arrays may be made up of circuit board epoxy fibreglass. Although these packages have reduced thermal mismatch and lower bond lead parasitics, their thermal conductance and hermeticity is relatively poor.

1.4.3 Surface Mounting

Surface mounting is a more advanced technology in which a circuit package is bonded to pads on the surface of a circuit board. Surface mount technology provides for smaller packaged devices which may be connected to both sides of the circuit board. As a result of these provisions, system performance is enhanced through reduced package and board parasitics.

From a manufacturing viewpoint, surface mounting is inherently easier than through-hole mounting though it does restrict test visibility of package pins and shows increased susceptibility to thermal stress through mechanically weaker connections.

Electrical connections between circuit pads and package are realised most often with wire bonding. Tape-automated bonding and flip-chip mounting are alternatives. The benefits and potential problems associated with each technique are addressed in the following sub-sections.

1.4.4 Wire Bonding

Using this technique, the integrated circuit is fixed to a lead frame with an adhesive which provides low thermal resistance. Aluminium wires then are individually attached so as to connect all circuit pads to the lead frame. The wire material is aluminium because of its low electrical resistance and malleability.

Although this technique so far has proved both cost-effective and reliable, it may not continue to be so as packages are required to integrate larger circuits with greater I/O requirements. Since the wires are individually bonded, there may not be adequate throughput in production.

Aside from these issues, typical bond wires exhibit a parasitic inductance of about 5nH. Will this property remain an acceptable parasitic as the levels of integration and synchronous circuit activity are increased ?

1.4.5 Tape-automated Bonding

Tape-automated bonding involves placing solder bumps on diced circuits which subsequently are aligned with copper leads fabricated in multilayer polyimide tapes. The bond is created by reflowing the solder and the tape is fed to automatic test and assembly machines which may place the device directly on to a circuit board.

This technique offers relatively high throughput since all bond connections are created simultaneously. In comparison with wire bonding, tape-automated bond lead separation is low due to the rigid supportive structure of the polyimide film. Tape-automated bonding requires relatively short leads and consequently exhibits reduced parasitic effects. The minimum bond lead pitch achievable with tape-automated bonding is about 0.08mm. The minimum pitch for wire bonding is about twice as much.

The disadvantage of tape-automated bonding is that the embedded bond lead patterns and pattern sizes are normally different for different circuits. This technique consequently is expensive and normally is used only for high-volume circuits.

A logical extension of tape-automated bonding is to make bond lead connections over the whole surface of the integrated circuit. This technique is known as area tape-automated bonding and its advantages are clear. Signals are no longer constrained to be at the periphery of the circuit.

Component	Capacitance (pF)	Inductance (nH)
68 pin plastic DIP ‡	4	35
68 pin ceramic DIP‡‡	7	20
68 pin SMT chip carrier lead ‡	2	7
68 pin PGA pin‡‡	2	7
256 pin PGA pin‡‡	5	15
Wire Bond	1	1
Solder Bump	0.5	0.1

‡ No ground plane; capacitance is dominated by wire-to-wire component.
‡‡ With ground plane; capacitance and inductance are determined by the distance between the lead frame and the ground plane, and the lead length.

Fig 1.13 Package Electrical Parasitics

Direct tape-automated bonding clearly is more suited than wire bonding, to large circuits with higher I/O requirements. Prototype direct tape-automated bonding on circuit boards has been developed to support more than 500 leads per circuit [b29]. Thermal mismatch between circuit and circuit board remains a problem for this technique which can but worsen as circuits become larger.

1.4.6 "Flip-chip" Mounting

Further reductions in bond lead length can be realised by placing solder bumps on the circuit, aligning them with contact pads on the package substrate, and reflowing the solder to create the bond. This method provides electrical connections with parasitic inductances of lower than 1nH and capacitances of lower than 1pF. In addition, the technique clearly offers the possibility of non-peripheral bonding. One obstacle facing this technique is associated with its resultant high thermal resistance. Unless a thermal contact is made to the back of the circuit, the thermal path from the circuit to the package is limited to the solder bumps.

Another obstacle is that of thermal mismatch between the circuit and the package substrate. This will lead inevitably to mechanical strain at each of the solder bumps and to unacceptably high failure rates. In large circuits especially, this strain is minimised by concentrating the bumps at the centre of the circuit [b30]. This constraint has the obvious effect that non-peripheral bonding would, in practice, be restricted to the centre of the circuit.

A summary of electrical parasitics associated with the above packaging techniques is given in Figure 1.13.

1.4.7 Thermal Properties

As integrated circuits become larger and operate at higher frequencies, they are bound to generate more heat. Since almost all failure mechanisms are enhanced by

higher temperatures, increasingly large portions of this heat must be removed if future packaged circuits are to have acceptable reliability. In order to meet commercial reliability requirements, circuits typically operate in the temperature range 0°C to 85°C and are stored in the range -55°C to 125°C.

In the case of plastic dual-in-line packages, there are three major thermal paths from the circuit to the ambient: from integrated circuit to package via the circuit bond material, from the circuit to the package pins via the wire bonds and bond lead frame and from circuit to ambient via the package plastic moulding. The second of these paths is often assisted by a metal sheet placed under the circuit.

The thermal resistance of a 40-pin dual-in-line package may be as low as 38°C/W using natural convection and 25°C/W with forced air convection. Given the desired temperature range for reliable operation, a dual-in-line package may dissipate up to 2W with natural convection and up to 3W with forced air convection.

Pin grid arrays, unlike dual-in-line packages, are always made up of a multilayer ceramic such as alumina or beryllia with a hermetically-sealed air cavity above. As was discussed in 1.4.2, cavity-down pin grid arrays may have heat sinks attached to one side. In these packages, the primary thermal path is via the circuit bond material, the ceramic substrate and the heat sink.

In the case of flip-chip mounting, the major thermal path from the circuit to the package substrate is via the solder balls and any attached heat sink. The scheme was developed by IBM for their 3081 processor unit [b31]. Heat is transferred from the circuit to pistons which, in turn, conduct it to a water-cooled plate. In addition, the circuit cavity is helium-filled to assist the cooling process. The resultant thermal resistance of the package is 11°C/W per integrated circuit site. Since the piston diameter is the same as the circuit length, circuit sites may be closely packed thereby minimising inter-circuit electrical parasitics.

The resultant package provides an environment in which densely packed integrated circuits may dissipate up to 4W of power and raise their junction temperatures only 44°C above that of the cooling water [b32].

1.5 Power Distribution

1.5.1 Simultaneous Switching Noise

Given that integrated circuit technology is being driven to facilitate larger circuits constituted by smaller devices operating at higher frequencies, it is important to consider the implications, if any, for necessarily global signals such as the circuit power supply.

It has been stated, in section 1.2 on interconnect modelling, that larger circuits will possess higher interconnect capacitance. In addition it is clear that, with increased switching speed and more densely packed devices, the rate of change of current (di/dt) needed to charge and discharge these larger capacitors similarly will be higher.

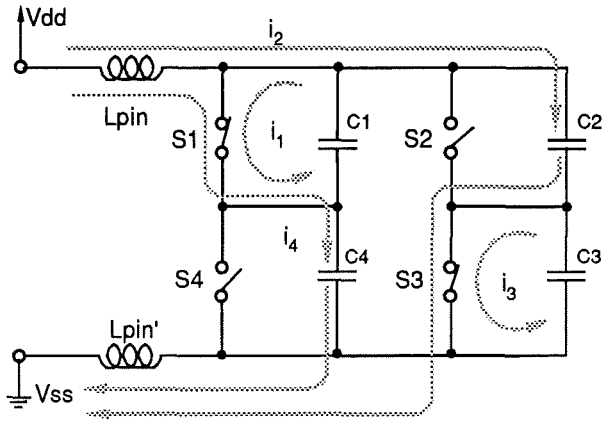
It is conceivable that di/dt will increase to levels such that the packaging- and interconnect-related inductive parasitics, discussed in sections 1.2 and 1.4, introduce power supply level fluctuations which significantly affect circuit performance. Such fluctuations are described by Faraday's Law written as

$$EMF = V_{disturb} = -Ldi/dt$$

At present, di/dt induced power distribution noise, is associated only with simultaneously switching circuit output drivers. As circuits are made larger and faster, or more synchronously active, power distribution noise associated with "core" circuitry may become equally important.

The complementary nature of digital CMOS circuitry is such that it is driven half passive. During a given cycle, the active devices will remove current from the positive supply and add current to the negative supply. This results in reduced power supply voltage integrity as the board power plane supplies current via the pin inductor.

Since the passive circuitry is connected to the relatively low inductance power supply



$$I_{int} = i_1 + i_2 + i_3 + i_4$$

$$I_{ext} = i_2 + i_4$$

Fig 1.14 Simplified Model of a CMOS Circuit, Package & Power Supply

distribution network via the on-resistance of the passive transistors, charge-sharing between active and passive circuits tends to compensate the demands of the active circuitry and thereby reduces the reduction in supply integrity.

It must be noted that the above effect acts only partially to compensate the current demands of active circuitry. Current always will flow in from and out to the external supply via the pin inductor. This fact can be explained with reference to the simplified electrical representation of active CMOS circuitry shown in Figure 1.14.

It is assumed that initially switches S1 and S3 are open and switches S2 and S4 are closed. In this state, capacitors C1 and C3 have a potential difference across them equal to the circuit power supply while capacitors C2 and C4 are discharged. It is next assumed that switches S1 and S3 close and S2 and S4 open. In this state, charge

redistribution associated with capacitors C1 and C3 will happen independently of the external supply whereas current will flow from the external positive supply via the pin inductor to capacitors C2 and C4. Similarly, current will flow from capacitors C2 and C4 to the external negative supply.

Even if, as in this simplified case, currents i_1 and i_2 exactly match i_3 and i_4, i_3 and i_4 will flow via the external pin inductor. In practical CMOS circuits, switches S1/S2 and S3/S4 do not precisely complement each other and the associated currents are seldom equal.

1.5.2 Noise Reduction

In section 1.2 it was established that if the signal source resistance R_S is significantly greater than the characteristic impedance of a transmission line, then transmission line phenomena become negligible. It therefore is reasonable to conclude that simultaneous switching noise will create minimum disturbance if the characteristic impedance of the power distribution network is kept to a minimum. Distribution network inductance must therefore be made as small as possible and network capacitance as large as possible.

Network inductance is interconnect- and package-related. Interconnect-related inductance does not, at present, contribute significantly to network inductance. As circuits become larger, however, its significance may increase and could be reduced by using thicker interconnect lines. Package-related inductance such as that associated with package pins and bonding structures may be reduced by increasing their cross-sectional areas, placing them closer to a ground plane and making them shorter. Inductance values for practical packaging options are listed in Figure 1.13 of section 1.4.

Once the separate inductance of individual package pins and bonding structures has been minimised, their effective inductance may be reduced further by placing as many as possible around the circuit thereby to provide parallel power supply current paths to and from the circuit and consequently to reduce absolute di/dt levels. Simultaneous switching noise is reduced.

The above technique has the added advantage of allowing isolation of especially noisy circuitry from relatively quiet high-precision circuitry.

In general, the effective inductance of the circuit power distribution network is proportional to the ground loop area. This may be kept small by forming a power circle around the circuit which then is multiply-bonded to the circuit. Ground loop area is small and therefore so is the effective inductance.

Network capacitance may be increased simply by connecting so-called decoupling capacitors directly to the circuit. If the capacitors are distributed evenly on the network, its characteristic impedance and consequent potential drop will appear to be reduced. A practical realisation of such decoupling techniques developed by IBM is reviewed in the following chapter.

It is important to point out that network decoupling capacitors are ineffective at

reducing noise associated with simultaneously switching output drivers. The "decoupling" path, shown in Figure 1.15, is contained within the circuit in the sense that a portion of the necessary supply current is provided to the decoupling capacitor by inactive circuitry via a low inductance path.

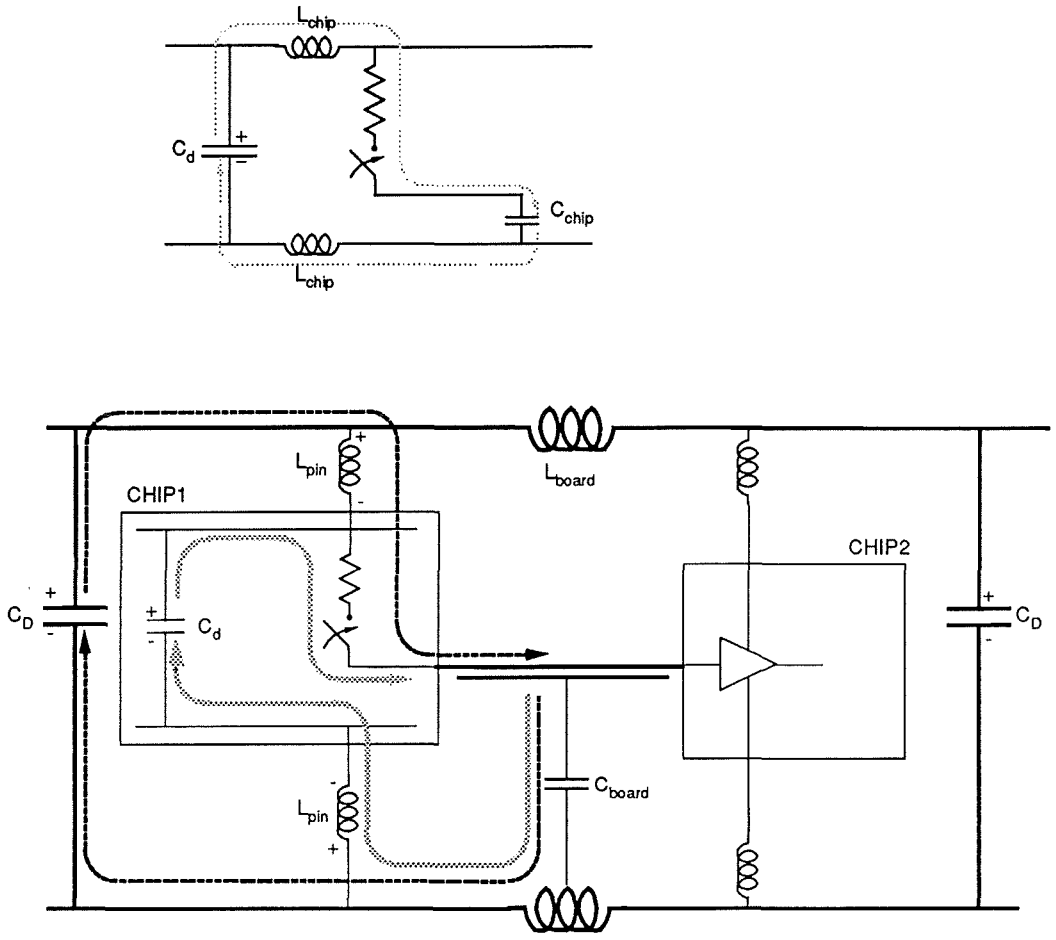


Fig 1.15 On-chip & Off-chip Supply Decoupling Paths

In the case of the output drivers, the path is not contained within the circuit. The system ground which is external to the circuit is connected to the circuit ground via the power supply pins. As can be seen from Figure 1.15, the ground loop is closed by a relatively high package pin inductance. A significant potential drop is developed across this inductor and as the output drives low and high the circuit supply will oscillate.

Network decoupling is useful only on the distribution network which serves "core" circuitry. For the reasons given, it is good policy to isolate core power supply from output driver supply.

Output buffers are often designed with more than adequate gain margin. Thus their contribution to power distribution noise is more than is necessary. It is possible to design output buffers whose gain is controllable so that switching speeds may be

reduced as much as the application will allow. Chapter two will discuss such di/dt-controlled output buffers and their effect on circuit performance overall.

1.6 Conclusions

It is clear that each of the issues addressed above will increasingly impact the level of synchronous activity which can be achieved with digital CMOS integrated circuitry.

To what extent will they impact the level of synchronous activity ?

The development of an answer to the above question, for the case of power distribution noise, forms the basis of this research programme.

The precise objectives of the programme are reiterated:

(i) To develop an accurate and flexible method of assessing the degree to which power distribution noise is generated in CMOS digital integrated circuits.

(ii) To determine the nature of the power distribution noise and its effect on the performance and electrical characteristics of the packaged integrated circuit.

(iii) To assess the extent to which power distribution noise limits the achievable level of synchronous circuit activity for CMOS digital integrated circuits.

Aside from a general understanding of the cause of power distribution noise, little is known of its nature or of the extent to which it is present in CMOS digital integrated circuits. More important, little is known of the extent to which power distribution noise may limit the performance of such circuits.

2 POWER DISTRIBUTION NOISE ANALYSES

2.1 Introduction

The objective of this chapter is to review all analyses of power distribution noise associated with highly synchronous CMOS digital integrated circuits.

At the time of writing, the sum total of published work on the subject is represented by only six papers. Of these [c01], [c02] and [c03] are concerned with simultaneous output driver switching and [c04], [c05] and [c06] have undertaken detailed analyses of power distribution noise associated with core sections of CMOS digital integrated circuits.

Of the papers concerning core-generated noise, [c04] was published in 1982 from Bell Laboratories, [c05] in 1987 from the Toshiba Corporation and [c06] collaboratively in 1990 from IBM Laboratories and General Technology Division. At the time of writing, these three are the only papers which discuss quantatively core-generated power distribution noise in CMOS digital integrated circuits.

2.2 The Delta-I Simultaneous Switching Problem

Ditlow and Brown [c01] refer to power distribution noise as the "delta-I simultaneous switching problem". They describe the problem from a system viewpoint and develop a Boolean model aimed at identifying potential simultaneous switching hazards. They confine their analysis to the effect of simultaneously switching output drivers.

In developing the model, they consider the definition of a set of input conditions which causes a corresponding set of output drivers to switch simultaneously. Since the output drivers must be preset to a known state, two sets of input conditions are required. Ditlow and Brown address the issue of finding a preset-set input sequence such that the number of simultaneously active outputs is less than a given technology limit. The limit is an assumed upper-bound beyond which the degree of simultaneous switching cannot be supported reliably by the power distribution network.

In order that the two-stage, preset-set, input sequence may be realised, the final form of their "delta-I hazard-flagging model" consists of a single output with twice as many inputs as the original combinational logic network.

The outputs of each model are input to a decision function used to assess the degree of simultaneous unidirectional switching; unidirectional since positive transitions will tend to cancel negative ones. The output of the decision function is then compared with a known technology limit and hence a delta-I switching hazard may be flagged. They conclude that the delta-I simultaneous switching problem can effectively be managed with the Boolean model.

Their analysis is of limited practical use since it is restricted in two ways. These are: (1) only the effect of output driver switching is considered; and, (2) the model rests on the fact that the constraining "technology limit" is known.

2.3 Ground Bounce Control in CMOS Integrated Circuits

Gabara and Thompson [c02] describe a technique for homogenising output driver rise/fall times so that high times, associated with poor processing, are reduced and low times are increased. This has the effect of equalising associated ground noise, or ground bounce, across the process window. They demonstrate the effectiveness of the technique by experiment.

The key feature of their technique is an integrated voltage source used to regulate the charge/discharge rate of the output driver. The voltage source is stable with respect to the process and generates a voltage according to its position within the process window so that the variation in output driver performance is minimised. Ground bounce is equalised with respect to all process variations.

Gabara and Thompson measured the delay, rise and fall-time of a conventional output driver and a voltage-controlled output driver fabricated in a standard 1.25 micron bulk CMOS technology. It is clear, from Figures 2.1 and 2.2, that the effect of the voltage-control is effectively to homogenise the process thereby effecting a reduction in the rate of change of supply current, di/dt, for the fast circuits and an increase for the slow circuits.

After simulating the effect of thirty-two synchronously active conventional and voltage-controlled output drivers, Gabara and Thompson claim a two-fold reduction in ground bounce resulting in a 10% reduction in driver delay.

Since if the ground bounce has been halved, the di/dt rate will have been similarly halved. With reference to Figures 2.1 and 2.2, it therefore is clear that, for the conventional drivers as compared with the voltage-controlled drivers, such a di/dt reduction will, on average, cause a 30% increase in delay.

For these numbers to be compatible with the authors' conclusions, it may be asserted that, everything else being equal, a two-fold reduction in ground bounce will cause a 40% reduction in driver delay.

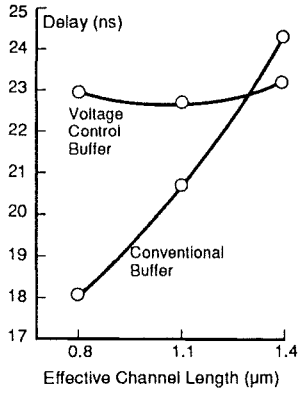


Fig 2.1 Delay vs Channel Length

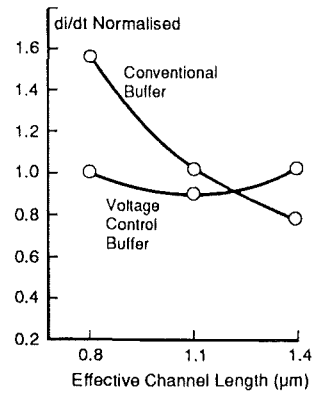


Fig 2.2 di/dt vs Channel Length

It is concluded that, in this case, delay is a strong function of ground bounce caused by simultaneous output driver switching.

In this case, also, the investigation is limited to the effect of simultaneously switching output drivers. Further, it may be asserted that di/dt limiting is not a solution to the problem of designing highly synchronous CMOS integrated circuits in as much as it is a method of reducing the degree of synchronous activity at the expense of circuit performance.

2.4 Power Bus Transients in Very High Speed Logic Systems

Ziesse, Werko, Dishman and Schlosser [c03] address the issue of providing a noise-free power supply to a system of high performance digital integrated circuits. They show that a traditional approach to power distribution and decoupling can result in significant supply perturbations.

Once again, the analysis is confined to the effect of simultaneously switching CMOS output drivers. The authors base their analysis on the equivalent circuit shown in Figure 2.3.

They represent each of thirty-two output drivers connected to a thirty-two bit bus by

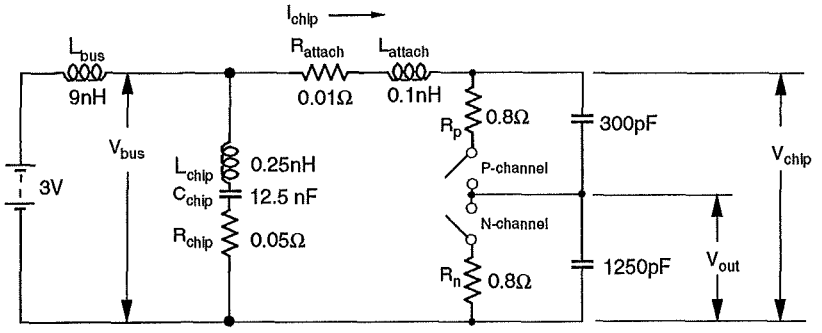


Fig 2.3 Equivalent Circuit

an n-channel and p-channel device configured as a single inverter with appropriate output loading. Channel resistances R_p and R_n are commensurate with one nanosecond rise-times.The 200pF capacitor represents n-well diffusion capacitances associated with p-channel transistors and the 1250pF capacitor the load associated with the logic buses.

The chip decoupling capacitor, C_{chip}, is assumed to be 12.5nF and further it is assumed that this value is attainable with a parasitic inductance of 0.25nH, L_{chip}. A 0.1nH inductor, L_{attach}, has been included to simulate the inductance of the bondwire and a 9nH inductor, Lbus, has been included to simulate a nine centimetre long circuit board wire. The authors assert that resistances R_{chip} and R_{attach} have been added to provide *appropriate loss*. They point out that the resonant frequency of L_{bus} and C_{chip} is around 15MHz and that, at this frequency, the power supply network impedance is 14 ohms. L_{chip} and L_{attach} are series resonant at

76MHz and above that frequency the network appears inductive.

Ziesse et al. exercise the simulation model for the case of driver inputs changing state at 5ns intervals thus corresponding to a frequency of 100MHz. This they term 100% switching to indicate a uniform unmodulated switching pattern. They then consider the case of three sets of six input transitions separated by 30ns intervals, 50% switching, and modulated at the resonant frequency of the power supply. Their results are shown in Figures 2.4 and 2.5.

While the authors acknowledge that the intentionally introduced 15MHz modulation of unspecified amplitude may not be representative of real systems, they note that it is responsible for a change in supply voltage which potentially is sufficient to cause logical errors in the same or neighbouring circuits.

They conclude that the simulation demonstrates two sources of power distribution noise. These are: (1) components whose frequencies are comparable with the system clock and its harmonics; and (2), low frequency components excited by specific logic patterns.

They recognise that low frequency modulations such as they intentionally introduced may be caused by package decoupling capacitance. They recognise that their low frequency modulation depends on the magnitude of the peak supply current at the resonant frequency of the power supply network and on the supply network impedance.

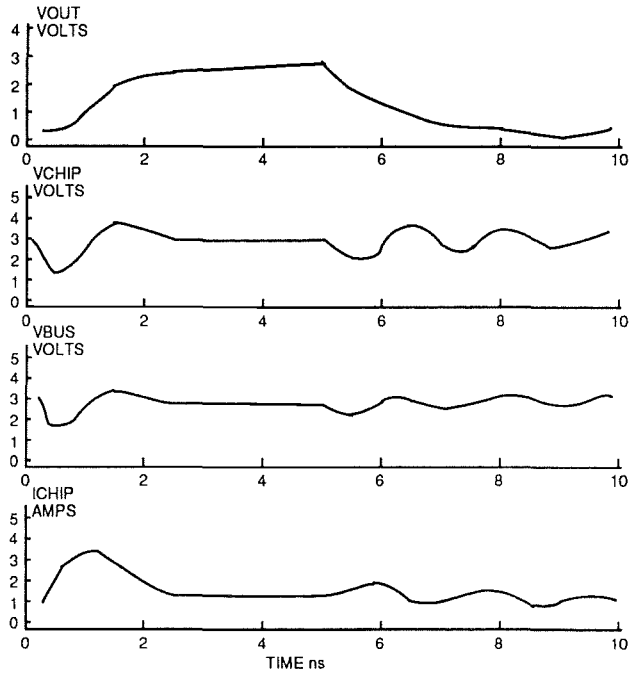


Fig 2.4 Computed Waveforms - High Speed CMOS Circuit Model

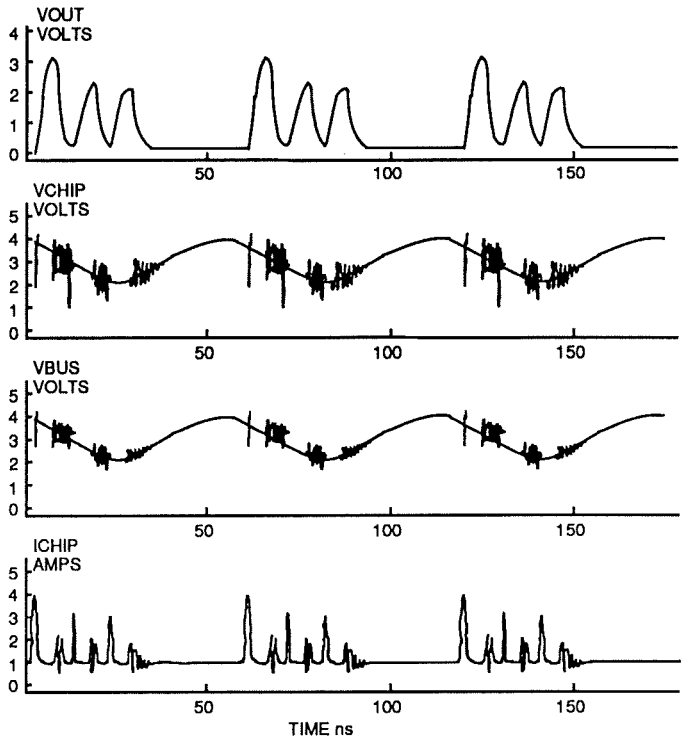


Fig 2.5 Computed Waveforms for 50% Switching showing Excitation of the 15MHz Power Bus Resonance

2.5 Noise Generation Analysis and Noise-Suppression Design Techniques

Shoji [c04] describes the analysis and design of the power supply for the Bell System's 32-bit CMOS VLSI microprocessor. The final structure of the power distribution network is shown in Figure 2.6. Shoji terminates each end of the vertical power bus with a bond pad which in effect reduces the power supply noise by a factor of four since current flow and impedance at connections to the power bus are halved.

In order to further reduce power supply noise to acceptable limits, Shoji separates the precharge timing of the PLA ROM from the decoder.

Shoji analyses *by simulation* the ground bus noise at many points on the distribution network. The results of this analysis, shown in Figure 2.7, illustrate clearly that noise due to PLA precharge is greatest at point A and lowest at point I and that noise due to I/O switching is significant at locations B and H.

It is noteworthy that the noise associated with I/O switching is significantly less than that of the core precharge circuitry focussed on in [c01], [c02] and [c03]..

Following Shoji's example Itoh, Nakagawa, Sakui, Horiguchi and Ogura [c05] report on a power supply noise model for megabit DRAM's. The model is used to predict and analyse power supply noise in the development of noise-suppressed DRAM circuits.

In order to predict and analyse quantitatively the peak switching current and associated power supply noise, their simulation model consists of three parts. These

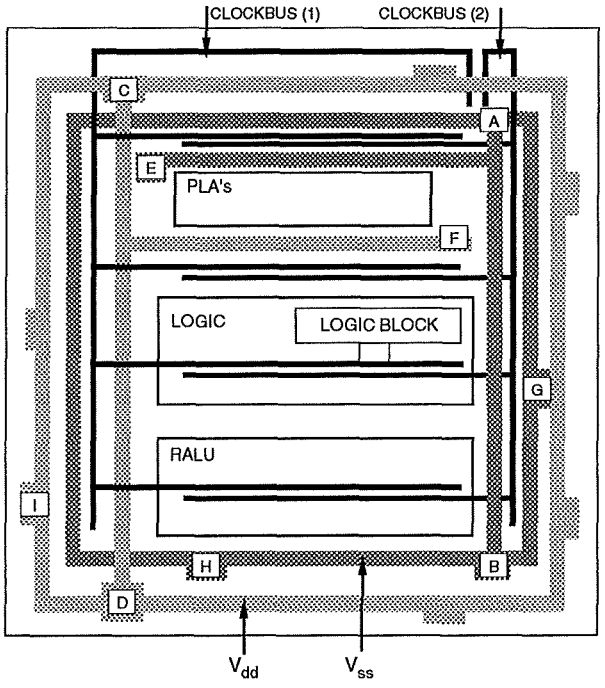


Fig 2.6 Clockbus & Powerbus Structure of BELLMAC-32A

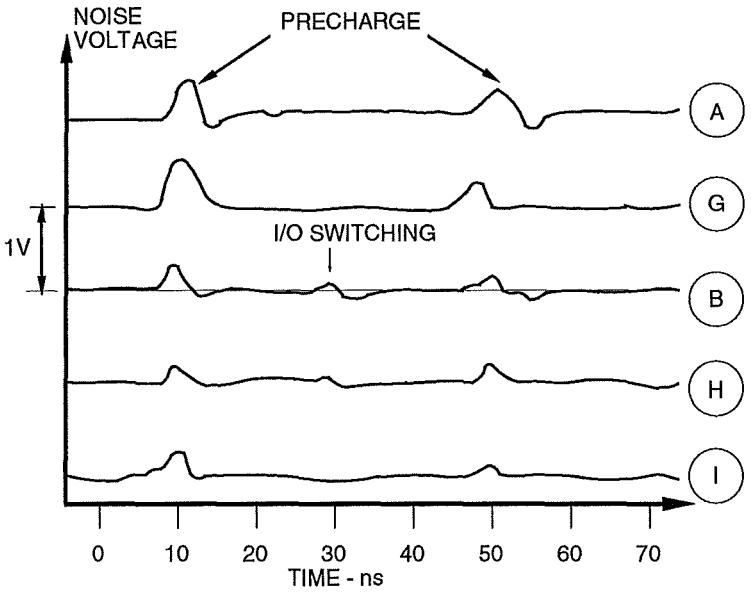


Fig 2.7 Noise Voltage vs Time

are: (1) parasitic impedances associated with the printed circuit board and the circuit package; (2) the DRAM cell and peripheral circuits; and (3) the appropriate output load. The emergent simulation model is shown in Figure 2.8. Itoh et al. have included a package-connected decoupling capacitor C_D in order to *stabilise* the power supply lines associated with the DRAM chip. In addition, they have included inductive and resistive elements of value 35nH and 0.5ohms to represent the combined electrical characteristics of the printed circuit board and the integrated circuit package. BL and BL-bar represent the true and complement bitline controls, C_B is the total bitline capacitance of some 400pF and Q1, Q2 and Q3 are bitline precharging devices. The

peripheral circuit block includes clock generators, address buffers, input buffers and decoders.

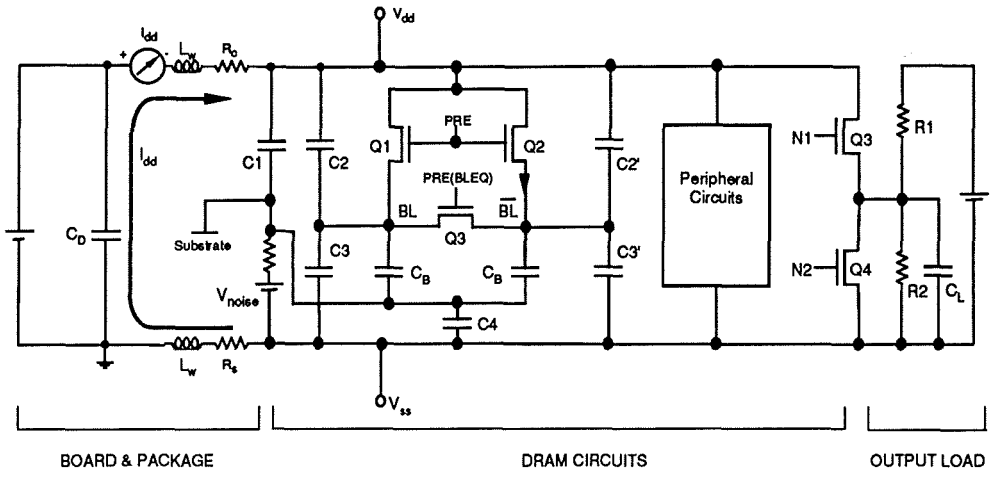


Fig 2.8 Simulation Model for DRAM Supply Noise

In analysing the current transients the authors assert that, since the substrate-bias generator has a high impedance output, it thereby causes the main current transient path to be associated with the peripheral circuit board and package parasitics. Other current transients participate through power supply, substrate and bitline coupling capacitors C1 to C4, C2' and C3'. Further, they assert that the power supply noise is dependent on a resistive and an inductive term, the resistive term depends on absolute values for peak current and the inductance term depends on the rate of change of supply current, di/dt. Their model predicts the inductive term to be dominant.

The authors have found that the above simulation model accurately predicts power supply noise for the DRAM device. They applied the same modelling technique to the design of a low-noise 1-Mbit DRAM in order that peak current and associated supply noise could be reduced without sacrificing DRAM access time.

In the new circuit, the degree of synchronous switching is reduced by separating the times at which the bitline precharge and equalising signals become active. The resulting reduction in peak current is shown in Figure 2.9. In addition, the authors separated these signals so that the negative-going edge of the precharge signal effectively is coincident with the positive-going edge of the equalising signal.

The authors conclude that the simulation model is an effective tool for high-density high-performance DRAM design and that its utilisation is essential for memory devices with high transient current demands.

The analysis technique presented by Itoh, Nakagawa, Sakui, Horiguchi and Ogura provides an accurate and detailed analysis of a real circuit which has been used in the development of a noise-suppressed high performance DRAM circuit.

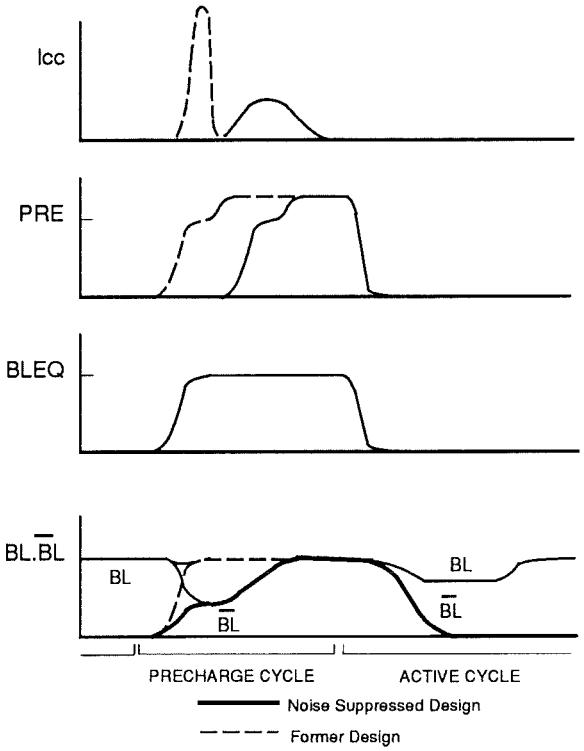


Fig 2.9 Operating Peak Current Icc & Control Signals for bit-line precharging circuitry

2.6 A CMOS Mainframe Processor with 0.5um Channel Length

Schettler, Haug, Getzlaff, Starke and Bhattacharyya [c06] describe the design of a CMOS VLSI processor chip set in terms of: logic design, device design, chip design and test methodology.

The processor is a development of mainframe processors from the IBM 4361 and the following 9370 system family. The objective of this new design was to exploit the fast improving cost and performance figures of advanced CMOS technology.

The chip set consists of five units: a fixed-point processor, an instruction processor, two caches and a floating-point coprocessor. The fixed-point and instruction processor are packaged along with the two caches on a multichip module. The floating-point coprocessor is packaged on a single-chip module. The associated transistor count is given in Figure 2.10.

Schettler et al. recognise that simultaneous circuit activity may cause an unacceptable degree of power distribution noise. In order to avoid this problem, they include on-chip decoupling capacitors between the positive (Vdd) and negative (Vss) supply lines. These capacitors occupy otherwise unused area below metal-2 Vdd and Vss supply lines and also unused cell area.

IP Instruction Processor	1,010k
FX Fixed Point Processor	188k
FL Floating Point Processor	417k
CA Cache	2 X 627k
Total	2869k

Fig 2.10 Transistor Count

The equivalent circuit for the on-chip decoupling capacitor is shown in Figure 2.11. The authors claim that the gate oxide of the n-channel and p-channel transistors provide a total decoupling capacitance of about 35nF. They assert that use of these on-chip capacitors has been predicted through simulation to reduce logic power distibution noise by a factor of 3 to 5. This has been verified by measurement. The results are shown in Figure 2.12.

The multichip module consists of a ceramic multilayer substrate with a brazed pin grid array. The substrate contains twenty signal and twelve power planes for signal shielding. One hundred and sixty power pins are distributed among the two hundred and fifty-six signal pins thereby to allow high switching rates.

Schettler, Haug, Getzlaff, Starke and Bhattacharyya do not give an account of how their simulations were implemented.

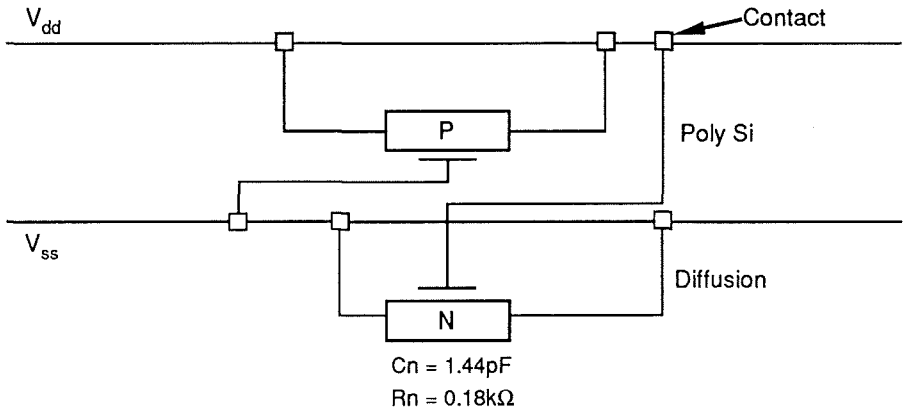


Fig 2.11 Layout & Circuit Schematic of on-chip decoupling capacitors

At the time of writing, this paper is the latest concerned with power distribution noise and indicates clearly that power distribution analysis and design must be adopted as part of the design of highly synchronous CMOS integrated circuits.

2.7 Conclusions

The analysis of Ditlow and Brown is of purely *academic* interest and would require a great deal of development for it to serve as a practical tool to be used in the design and analysis of integrated circuits.

Gabara and Thompson give an early account of a technique which is used widely to

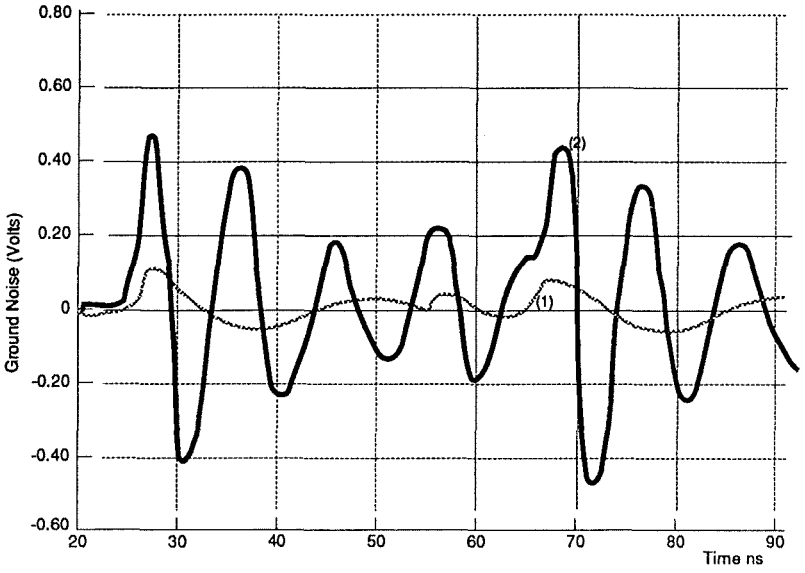


Fig 2.12 Chip Ground Noise with (1) and without (2) on-chip decoupling capacitors

reduce power distribution noise by alleviating over-capacity in areas of the circuit whose performance exceeds that of the local rate-determining circuit.

Since this method depends on tapping off localised performance over-capacity, it clearly is not a long term solution to the problem of distributing power to increasingly synchronous integrated circuits.

The point to be made from this paper is the fact that Gabara and Thompson conclude a strong dependence of circuit performance on ground bounce.

Ziesse et al. present a simulation model which may be used to predict power distribution noise associated with simultaneous switching of output drivers. They conclude these changes may cause logical errors.

The paper by Shoji and by Itoh et al. demonstrates the effectiveness of accurately assessing, by way of full circuit simulation, the amount of power distribution noise during the design and development of high performance CMOS integrated circuits. They conclude that the inductive noise component is more significant than the resistive.

These papers are the first to consider in detail the effect of core circuitry on power distribution noise.

The final paper by Schettler et al. provides an account of how power distribution noise was recognised as a potential problem, estimated through simulation and reduced through power supply decoupling. They achieve a reduction in power distribution noise by a factor of 3 to 5 by increasing network capacitance to around 35nF.

This recent paper, published in October 1990, illustrates the growing awareness of

the need accurately to estimate power distribution noise in the development of highly synchronous CMOS digital integrated circuits.

As was the case for integrated circuit testability about a decade ago, it is becoming clear that an ad-hoc approach to power distribution is no longer adequate. New and more sophisticated methods of noise assessment and associated power distribution performance should be developed.

The development of such a method for the case of an *inherently* highly synchronous array-based architecture, the systolic array, is described in the following chapter.

3 POWER DISTRIBUTION NOISE IN AN ARRAY-BASED ARCHITECTURE

3.1 Introduction

The objectives of this chapter are (1), to introduce the concept of systolic algorithms and architectures as used to address highly compute-bound problem-solving in digital signal processing; (2), to develop a simulation model that can be used to assess power distribution noise associated with systolic array integrated circuits of ever-increasing size; and (3), to derive, from the results of the simulation model, the extent to which power distribution noise will limit the achievable level of integration and performance of integrated circuits based on such highly synchronous architectures.

In order to explain clearly how the above objectives were achieved, the following structure has been adopted for this chapter:

Section 3.2 is an introduction to the principle of systolic algorithms and architectures and includes three examples of systolic array integrated circuits that have been implemented to undertake specific compute-bound tasks with application in digital signal processing.

Section 3.3 is a description of the development of a noise model for use in predicting power distribution noise and associated voltage integrity levels for a generic systolic array integrated circuit requiring 1,020,000 devices.

Section 3.4 will describe a method by which the voltage integrity levels, predicted by the simulation model, can be used to derive the associated performance implications of systolic array integrated circuits. Since such implications are inextricably linked to the power distribution technology, they are derived for both standard and non-standard, but emerging, power distribution technologies.

Section 3.5 will address the sensitivity, of the simulation model, to each of the assumptions and design choices instituted therein. It is not difficult to imagine that a different simulation model may produce significantly different results.

Section 3.6 will draw together and summarise each of the conclusions for sections 3.3 to 3.5.

3.2 Architectural Overview

3.2.1 Systolic Architectures: The Basic Principle

A systolic system consists of a set of interconnected cells, each capable of performing some simple operation. In terms of design and implementation, it is clear that simple and regular communication and control structures have a distinct advantage over complex and less regular ones and it is for this reason that cells, in a typical systolic

system, are interconnected to form a systolic array or a systolic tree. Information in a systolic system flows between cells in a pipelined fashion and communication with the outside world occurs only at the boundary cells. A systolic array is a two dimensional pipeline processor.

Computational tasks may be classified into two families: compute-bound computations and I/O-bound computations. If the total number of operations, in any computation, is larger than the total number of input and output events, then the computation is compute-bound. Otherwise, it is I/O-bound. The ordinary matrix-matrix multiplication algorithm, for example, represents a compute-bound task since each matrix element must be multiplied by all elements in each row or column of the other matrix. Adding two matrices, on the other hand, is I/O-bound since the total number of additions is not larger than the total number of elements in the two matrices. Clearly, any attempt to speed up an I/O-bound computation must rely on an increase in memory bandwidth which can be realised either by the use of fast components, which could be expensive, or the use of interleaved memories, which could create complicated memory management problems. Speeding up a compute-bound computation, however, may be accomplished, in many cases, in a relatively simple and inexpensive manner through the use of systolic algorithms and architectures.

The basic principle of a systolic architecture, a systolic array in particular, is

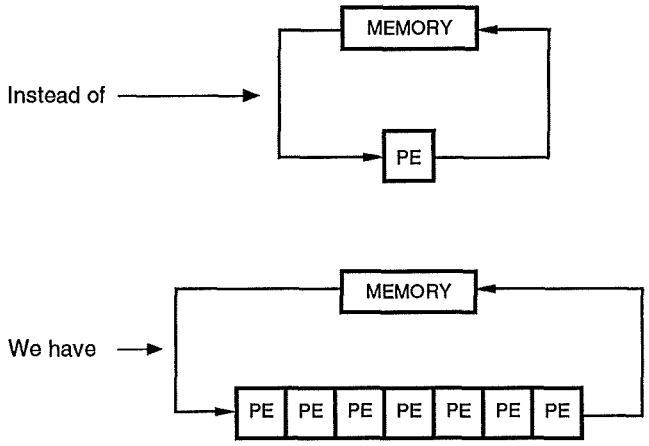


Fig 3.1 Basic Principle of Systolic Array

illustrated in Figure 3.1; by replacing a single processing element with an array of smaller processing elements, a higher computational throughput can be achieved without increasing memory bandwidth. The function of the memory in the figure is to "pump" data through the array of cells in a manner that is comparable with the biological heart.

The crux of the above approach to computation is to ensure that as soon as a data item is accessed from the memory, it is able to be used effectively at each cell, that it passes, while simultaneously being pumped from cell to cell along the array. It is possible to enforce such a data flow scenario in a wide range of compute-bound tasks in which multiple operations are performed on each data item in a repetitive manner [d1].

The ability to use each input data item repetitively, thereby achieving high computational throughput with only modest memory bandwidth, is just one of the many advantages of the systolic approach to computation. Other advantages include modular expandability, simple and regular data and control flows, the use of relatively simple and uniform processor cells and the elimination of global data broadcasting.

3.2.2 A Bit-level Systolic Array for Matrix-vector Multiplication

A general n-point matrix x vector transform can be expressed in the form Ax=y, where A is an n x n matrix, x is an n-element vector of data values and y is an n-element output vector. In general, the elements of A will be multi-bit words, but it may be assumed that each can take only the values 0 or 1 and therefore can be written in 1-bit form. A general m-bit transform then can be implemented using m bitslices in parallel. McCanny and McWhirter [d02] propose a systolic array circuit for the pipelined computation of this matrix x vector transform and their circuit, for the case of a four-point bit-slice transform, is shown in Figure 3.2; heavy dots have been used to represent latches and open circles denote the basic processing cells.

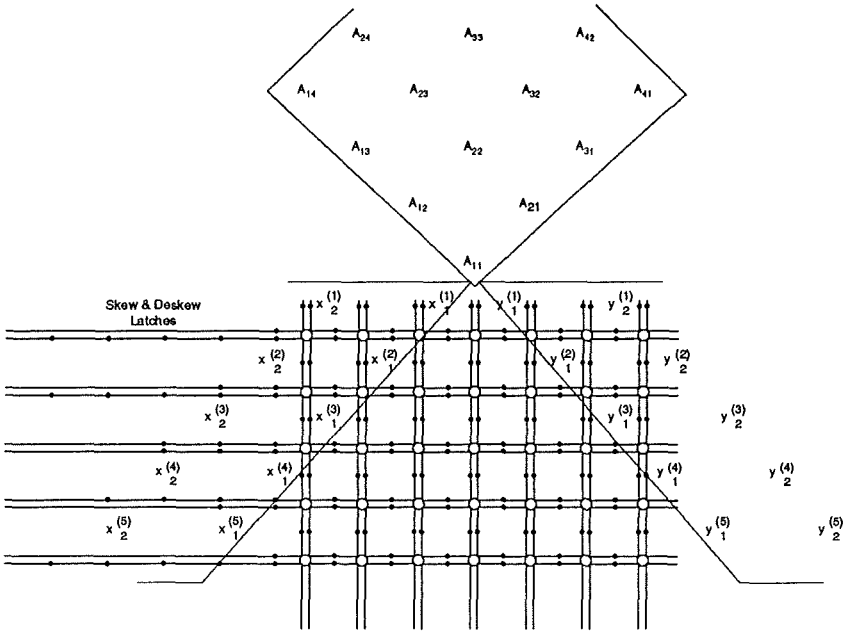


Fig 3.2 Four-point Bitslice Transform

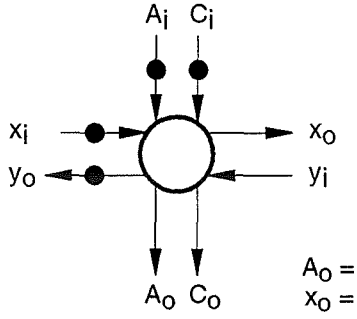


Fig 3.3
Constituent Processor Cell Logic Function

$A_O = A_i$
$x_O = x_i$
$y_O = y_i \oplus C_i \oplus (x_i.A_i)$
$C_O = y_iC_i \oplus y_i(x_i.A_i) \oplus C_i(x_i.A_i)$

The circuit consists of an array of latched gated full adder cells interconnected on an orthogonal lattice with connections between nearest neighbour cells only. The logic function of a constituent processor cell is as shown in Figure 3.3.

During each cycle of the system clock, each cell receives input bits A_i, C_i, x_i and y_i from neighbouring cells. A represents an element in the matrix A_i, x_i is one bit of an incoming data word which has to be transformed, y_i is one bit of the accumulating result and C_i represents a carry bit from a previous stage in the calculation. The resulting sum bit y_o and carry bit C_o subsequently are latched into neighbouring cells on the next clock cycle; at the same time, bits A_o and x_o similarly are latched into the neighbouring cell but without being altered in value.

The overall operation of the array is as follows: data words, x_i, are input to the array from the left, during every second clock cycle; successive bits are staggered by inputting the parallel words through an arrangement of skew latches and the least significant bit, $x_i(1)$, enters the array one clock cycle ahead of the next sigificant bit. The individual bits move one bit to the right during each clock cycle.

For the example in Figure 3.2, it is assumed that the incoming data are three-bit two's complement numbers which, before entering the array, have to be sign-extended to five-bit two's complement form so as to accommodate the range of the answer. A total of seven by five processing cells then are required, i.e. 2n-1 x l cells, where l is the number of bits in the result produced by the bitslice array. Bits representing the elements in the matrix A are organised so that they move vertically down the array as shown in Figure 3.2. On entering the array, the output words y_i are initialised to zero and move from right to left, similarly with their bits staggered. This means that the kth bit of a word, $y_i(k)$, meets all of the terms that are required to form the sum the product $A_{ij}x_j(k)$ being formed by typical ANDing. Any carry bits which are generated in the course of this summation are latched vertically downwards. The carry-save principle is utilised, therefore, and this is the reason for the stagger on each bit of the x_i and y_i words.

Having traversed the array, the output words y_i are transformed completely and one bit of each significance level, $y_i(k)$, emerges on every *second* clock cycle. The resulting output word can be deskewed by passing it through another arrangement of skew latches as shown in Figure 3.2.

3.2.3 A Systolic Array for Correlation

Christie [d03] describes a sixty-four stage bitslice digital correlator that makes use of systolic algorithms to provide a parallel correlation rate of twenty million samples per second. The design is for four-bit data and one-bit reference and allows for modular expansion of correlation length, reference width and data width, without recourse to additional control logic.

The design computes the correlation y(k) of sixty-four reference coefficients a(i) with a data sequence x(k+i). Internal operations are performed in two's complement or positive magnitude; the result is in two's complement and may be compared to a sixteen-bit threshold.

Without the use of systolic techniques, the implementation of a sixty-four stage correlation function would require a sixty-four stage shift register, sixty-four multipliers and a sixty-four input summing circuit. This systolic approach makes use of multiple adders to accumulate the result over a sample window and is capable of producing the final result as a parallel sixteen-bit word.

By reducing the coefficient length to one-bit, Christie is able to reduce the multiply function to a relatively simple gating function and, in so doing, the systolic processor emerges as a gated full adder. If the adders are then made to operate synchronously,

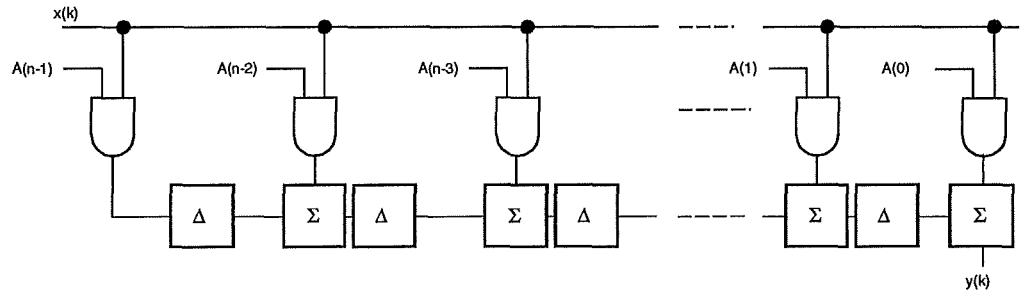


Fig 3.4 Equivalent Architecture of Systolic Correlator

the shift register delay is, in effect, transferred from the data stream to the results stream thereby yielding the equivalent architecture shown in Figure 3.4.

Christie has implemented this architecture as an array of 640 gated full adders arranged as sixty-four by ten-bit pipelined bit-serial accumulators.

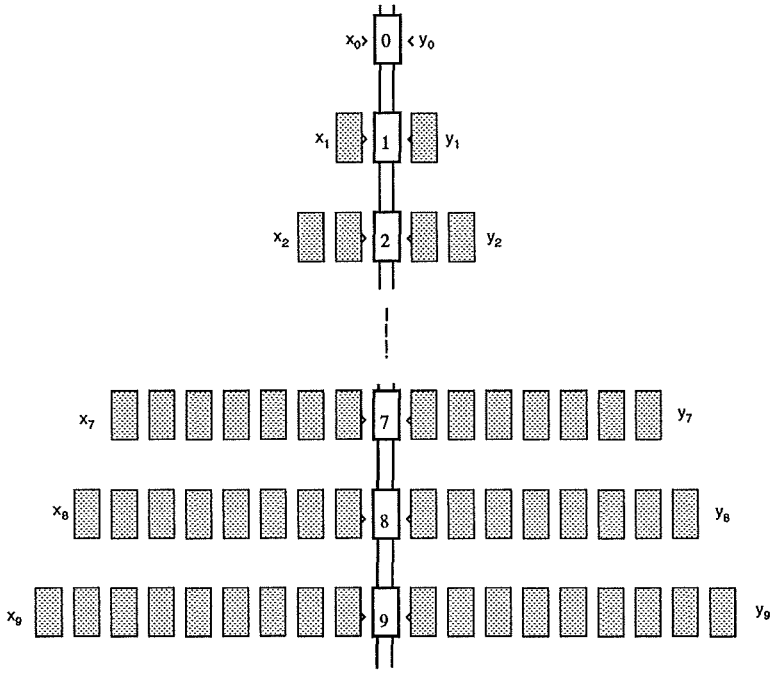


Fig 3.5 Bit-Serial Operation of Systolic Correlator

Taking one of the sixty-four columns, Figure 3.5 shows the bit-serial operation associated with each of the sixty-four ten-bit accumulators; the total result is calculated at the rate of one bit per half clock cycle from the least significant bit downwards, with the carry propagating down the column as each bit is calculated. This operation results in the accumulated result becoming available for further serial processing with one half cycle delay between successive bits.

The input data must also be presented in this way in order that they coincide correctly with the accumulating total, thereby allowing columns to be connected directly together, so as to accumulate the complete sixty-four-stage result in a pipelined operation. The data word is skewed by one half cycle per bit before being input to the array and the result is deskewed on exit from the array. The architecture may be used to correlate a continuous four-bit data stream with a one-bit reference over sixty-four stages.

The sixteen-bit reference data are loaded into the correlation array through serial shift register stages which are clocked on *alternate* phases thereby resulting in a total delay of thirty-two clock cycles. Since half cycle stages are used, and the reference data are presented serially at the rate of one-bit per clock cycle, each reference bit is held across two shift register stages. These shift register stages are referred to as the *holding register.*

The fact that the four-bit data stream passes across the correlation array in step with the reference data across their holding register, the relative timing for reference data loading is simple. After thirty-two cycles, as the first bit of the data stream reaches the end of the array, the first bit of the reference data reaches the end of the holding register whereupon it is gated into the array by the enable bit so as to calculate the first bit-product.

In addition to the basic correlation circuit, peripheral circuits are included to allow expansion in correlation length, reference width and data width. The circuit has been designed as a generic building block for many correlation functions. Typical examples are spread spectrum radio and radar and sonar pulse compression and detection.

3.2.4 An "efficient" Systolic Array for Distance Computation

Catthoor and de Man [d04] propose a systolic array for the implementation of "efficient" distance computation algorithms as applied to a video-codec system proposed by Verbiest et al.[d05].

The kernel of the distance computation algorithm lies in computing the distance between a block B(x,y,i) that is centred around the coordinates (x,y) in a frame i and the possible previous locations B(x-dx,y-dy,i-1). The following formula, proposed by Verbiest, is invoked as a simple and sufficiently accurate distance measure:

$$\text{Distance} = \Sigma_k \; | B_k(x,y,i) - B_k(x-dx,y-dy,i-1) |$$

The summation index 'k' ranges over the entire image block and in order to derive the desired motion vector, the derived distances are compared to determine the minimum.

To be capable of processing real-time video signals, Catthoor and de Man predict a required pixel rate of 10.4MHz and further they calculate that, with the above distance measure, around five billion additions or subtractions have to be performed during each second. It is for this reason that the authors have opted for a systolic array approach.

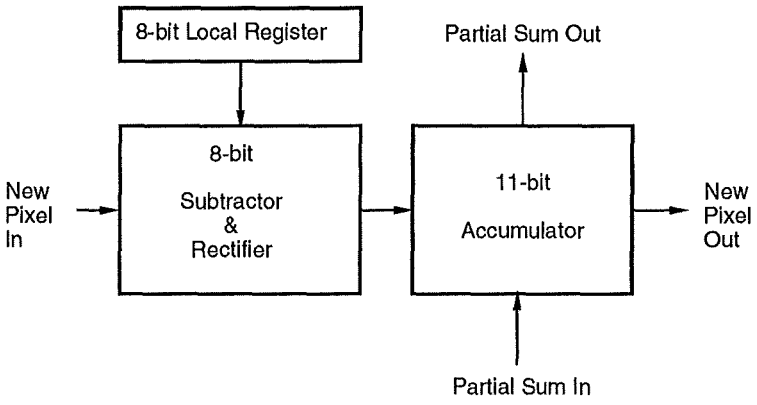


Fig 3.6 Cell for Motion Detector

In mapping the above distance measure algorithm on to a systolic architecture, Catthoor and de Man have extracted from it a common set of operations; these turn out to be the computation of the absolute value of the difference, between the block coordinates, combined with an addition to the accumulated sum. Catthoor and de Man assert that it is evident that each systolic array processing element should be constituted by a subtractor, a "rectifier" and an accumulator configured as shown in Figure 3.6. In addition, they assert that since the reference window will remain fixed until all comparisons have been performed, the pixel values of the previous time frame may be stored locally in the cells where, occasionally, they can be updated.

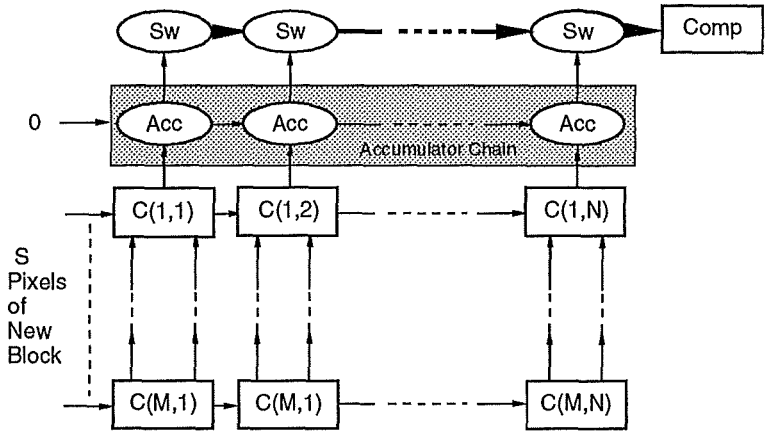


Fig 3.7 Matrix of Processor Cells

When the incoming pixel rows of the block B(x,y,i) are chosen to be incoming from left to right, the accumulation should proceed in the vertical direction as indicated in Figure 3.7. In principle, all terms of the distance relation are computed

simultaneously in the array cells and consequently have to be summed over the respective columns. This process will occur if an accumulator chain is added, at the top of the systolic array as shown in Figure 3.7.

Finally, the distance values have to be transferred to a comparator unit and, for this purpose, a row of two-way multiplexors is provided; these have to tap successive accumulators that contain the final results and then send them to the comparator.

The arriving block next has to be compared with the reference blocks that have been shifted downwards in the array. For this purpose, the data block can be repeated so that successive data blocks may be shifted one cell distance thereby resulting in the input data flow shown in Figure 3.8. Summation of the individual and distinct column sums is performed in the same additional accumulator chain for all successive,

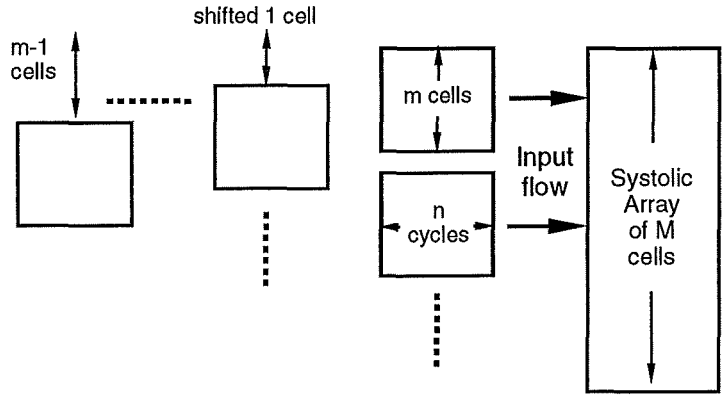


Fig 3.8 Input Data Flow

vertically-shifted, blocks in Figure 3.8. To achieve this, Catthoor and de Man have implemented two vertical data flow paths: one for the partial results and one for the final column summations; at the appropriate time, when a partial sum has been completed, these two busses are shorted by means of a switch.

Finally, in order to obtain correct accumulation of the individual columns, the partial sum inputs have to be initialised to zero. This operation easily can be performed by selecting directly the output of the subtractor as the new partial sum instead of the incoming partial sum.

Catthoor and de Man present a systolic array architecture to compute the motion vectors for image blocks in consecutive time frames over a restricted window. All operations are performed in a fully pipelined fashion resulting in real-time behaviour with a pixel rate of 10.4MHz. The data are updated dynamically and the control signals are derived from a set of only four master signals. Moreover, the I/O requirements are handled on-chip resulting in realistic off-chip communication with a limited number of pins.

3.2.5 Conclusions

With reference to the above architectural overview, it is clear that systolic array solutions to digital signal processing problems consist typically of large arrays of simple, often bit-level, processors being controlled to operate synchronously by circuitry that is usually resident at the periphery of the circuit. In addition, it may be concluded that, normally, their regular communication and control structure allows the design of the associated control circuitry to be simple and consequently relatively small. In each of the cases reviewed above, the constituent processor has consisted of a full adder along with some latches and, in this sense, each is typical.

In addition, the examples have illustrated that it is not usual for all the constituent processors to become active synchronously and that, in general, the degree of synchrony is dependent on the "efficiency" of the systolic algorithm and associated architecture Although algorithms and architectures can sometimes be improved upon [d06], it is common, in systolic array circuits, for each constituent processor to become active on each alternate phase of the system clock; this is the case in each of the examples that were reviewed.

Finally, it is concluded that, the fact that systolic arrays possess short interconnections that permit high communication bandwidths and minimum delays means that circuits based on a systolic approach are capable of performances in the region of many tens of megahertz.

This fact, in combination with the fact that their inherently regular structure lends itself to yield enhancement methods results in an integrated circuit that is extremely synchronous. Power distribution noise-related performance limitations are expected to reveal themselves first with such architectures.

3.3 The Development of a Simulation Model

3.3.1 Overview of Requirements

The objectives associated with the development of the simulation model are: (1) to develop a simulation method appropriate for the transient analysis of global power distribution networks in integrated circuits; (2) to use the simulation model to assess the magnitude and nature of power distribution noise; and (3) to derive from the noise predictions the extent to which circuit performance is limited by power distribution technology.

The systolic array processors described in section 3.2 are typical in that the constituent array processor cells are gated full adders controlled by an *alternate* phase clocking scheme and that the cell array constitutes a much larger area than the surrounding control and interface parts. This second property forms the basis of the first assumption implicit in the simulation model. It is assumed that only the array processors consume power.

Given that the first objective of this model is to develop a simulation technique that will allow *transient* analyses, accurate circuit timing information is an essential ingredient of the model. The precise level of detail that is needed will emerge during the development of the model and during the analyses of results at the end. Since, at this stage, little is known about the the sensitivity of current transients to circuit timings, it was decided to begin with the effects of clock skew.

With reference to the examples in section 3.2, it is clear that systolic array integrated circuits depend, for their timing, on an alternate-phase clocking scheme derived from a single source. In order to avoid exposing this clock too often to the lossy transmission lines fabricated on integrated circuits, it is envisaged that a locally synchronous distribution scheme, such as described in section 1.3.2, be used.

Such schemes may broadcast the clock *globally* at a given frequency and then redefine and resource *locally* at higher frequencies. Although, locally synchronous schemes which are *frequency-hierarchical* can better reduce the overall clock skew when compared to single-frequency schemes, it was decided to use a single-frequency scheme. The justification for this choice is that it is intended to assess the extent to which clock skew, in both its global and local forms, affects power supply current transients. The adoption of a frequency-hierarchical distribution scheme would only add a redundant degree of freedom to the analysis in as much as it would not provide any more useful information with regard to the sensitivity of current transients to clock skew but only to what extent different levels of frequency-hierarchy may reduce the overall degree of clock skew.

A single frequency locally synchronous scheme can be separated into its global and its local components. The inherent high regularity of the systolic array processor suggested a tree-based global component with each tree branch porting to an area of circuitry that is about equivalent to a *LSI-sized* array processor. This choice was made to take advantage of established clock distribution techniques.

In order for a local distribution scheme to be designed, the number of processor cells in the *LSI-sized* array needs to be decided. It is important that a realistic number is chosen, since in effect, the larger the number, the coarser is the information with regard to global clock skew effects and the more difficult is the task of local clock distribution. The numbers which were chosen for the array are for those of the systolic correlator circuit, as described in section 3.2.3, but with some redundant rows and columns. At the local or LSI level, the array is assumed to be eleven processors high and eighty processors wide.

The gated full adder used in the systolic correlator and shown in Figure 3.4 was taken as the constituent processor cell. In order that the physical size of this cell could be approximately ascertained, the cell was designed and a mask layout was generated. The cell layout was based on two micron fabrication technology and emerged at around 300 microns high and around 60 microns wide thereby implying a local array size of around 3.3mm high and around 5.0mm wide. Each cell requires sixty-eight transistors implying a total of around 60,000 transistors in each *LSI-sized* array.

In order to evaluate different methods of local clock distribution, the above information must be used in combination with realistic values for the resistance, capacitance and inductance per unit length that is appropriate for transmission lines on integrated circuits. This information was generated using the microstrip analysis programme MCLINES using dimensions and dielectric constants appropriate for two

micron wide metal interconnect on a two-micron bulk CMOS process.

As suggested by Antinone and Brown [b09], good simulation accuracy is obtained if each *section of interest* is modelled using an RCL ladder network whose number of elements is such that the time constant of each element is one tenth of the lowest time constant of interest. In this case, therefore, if each *section of interest* is one column of array processor cells, then a ten-element ladder network should be used to describe this section and thus an RCL ladder element for each 0.33 millimetre of track is needed.

Several schemes were evaluated for local clock distribution to each of the processors in the array of eleven processors high and eighty processors wide. The scheme shown in Figure 3.9 emerged as providing least overall clock skew. This scheme was adopted for local distribution to the array.

The effects of *global* clock skew were addressed similarly using the model shown in Figure 3.10. In this case, each section of interest is one processor array width so that an RCL ladder element for each 0.5 millimetres of track is needed. The degree of skew present for up to seventeen arrays, equivalent to a circuit of around eight centimetres in length, was determined. The resultant skew is shown in Figure 3.11.

In the design of the driver stages, the following had to be borne in mind: (1), with CMOS technology, due to its complementary nature, the use of multi-stage drivers is the most appropriate method of achieving a high gain digital clock driver; (2), transistor gate widths, in successive stages of a multi-stage driver, should be around three times those of the preceding stage [d07]; and (3), higher gain stages have, associated with them, higher power supply current transients and lower levels of clock skew until the capacitative effects of the driver stages themselves approach those of the line that it is being used to drive. In short, a technological drive-limit is arrived at.

In this case, four-stage drivers were chosen so as to provide good gain with non-inverted outputs; transistor widths were chosen so that each successive stage in the four stage device was three times wider than the preceding stage and so that the point was just reached at which further increases in the gain of each stage did not appreciably improve the overall level of clock skew.

3.3.3 Current Flow Modelling

With both the global and local components of the clock distribution scheme defined, information concerning the current flow in the supply rails of the constituent processor cell was ascertained. This was done using SPICE circuit simulator [d08] with transistor models representing two micron bulk CMOS. Test vectors were chosen to represent 20MHz clock rates with 2ns transitions. In addition they were chosen so that all inputs switched simultaneously from logic low to high thereby effecting maximum cell activity.

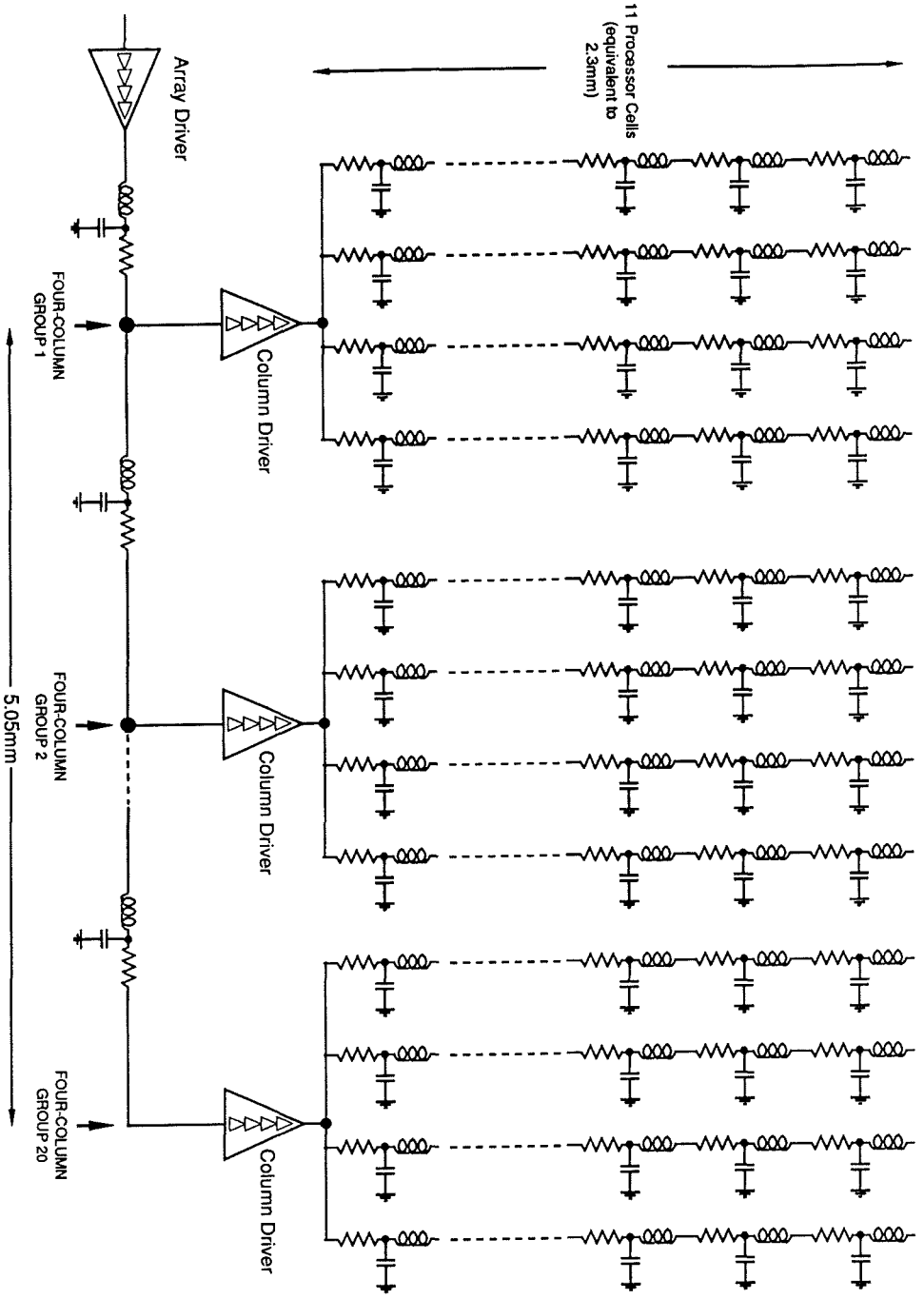


Fig 3.9 Local Clock Distribution Network

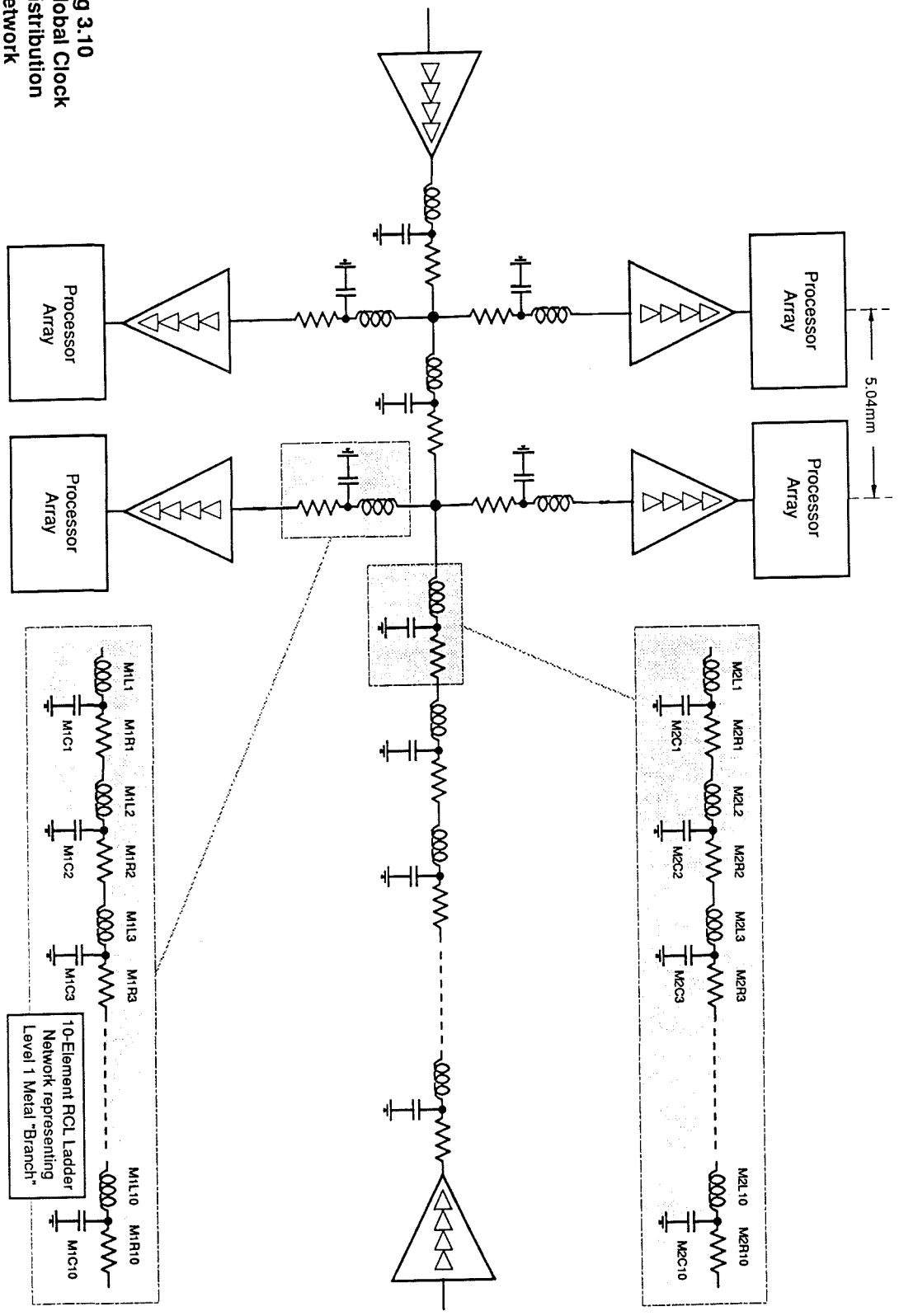


Fig 3.10 Global Clock Distribution Network

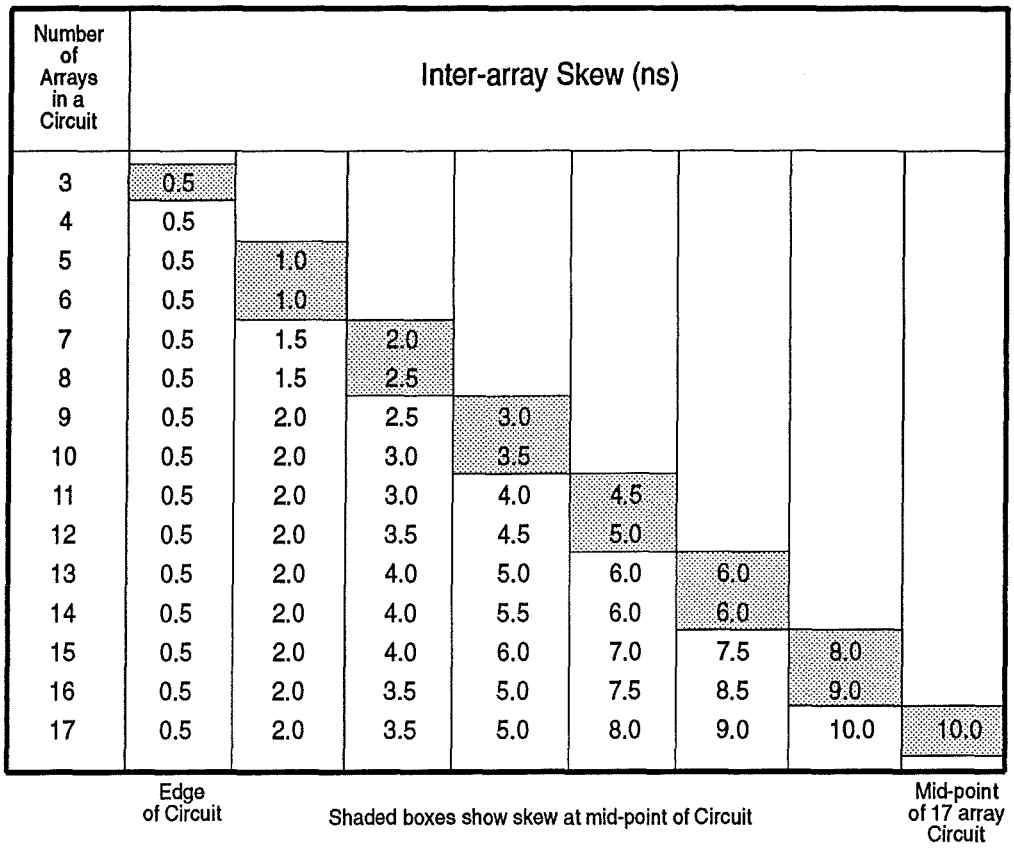


Figure 3.11 Global Clock Skew vs Array Size

Supply rail current flow for the gated full adder processor cell is shown in Figure 3.12. It is notable that almost all current flow is associated with the rising and falling edges of the clock. An identifying feature of CMOS technology is that, with the exception of leakage currents, it does not dissipate quiescently. The computational complexity of SPICE simulation algorithms is dependent on the number of active components raised to the power of 1.4 so that circuit simulation of many systolic processor arrays is computationally-bound. Circuit simulation of an entire array of these cells, though computationally demanding, is possible with modern computing resources such as the DEC VAX-780 series.

To assess the combined effect of many such arrays switching simultaneously, an equivalent circuit representing the current flow in each of the positive (Vdd) and negative (Vss) power supply lines should be designed. The requirements of this equivalent circuit are: (1) it should accurately mimic the current flow; and (2), it should be made up of as few active components as possible in order to minimise computational complexity.

Figure 3.13 illustrates the adopted solution. A voltage-controlled current source with, as control, a piecewise-linear voltage source having a profile that is proportional to that of the simulated current flow in each of the Vdd and Vss power supply lines of the processor cell. Notice that the voltage-controlled current sources for each of Vdd and Vss have opposite polarities.

This technique meets the requirements of the equivalent circuit; the current flow profile can be duplicated in as much detail as is required and the circuit contains only one active component thereby reducing the required computing complexity by 139; the number of transistors in each processor cell raised to the power of 1.4 divided by the number of active components in the equivalent circuits for both the Vdd and Vss supply current raised to the same power. In addition, the technique has the advantage of being simple. The current flow, as predicted from the simulation of the actual processor cell, can be compared to that of the equivalent circuit in Figure 3.14.

Figure 3.12 Supply Current vs Time

Although good agreement between equivalent circuit and full circuit simulation is shown by Figure 3.14, the circuit is only equivalent for the case of a power supply whose positive and negative rails are held fixed at five and zero volts respectively. The complementary nature of CMOS technology is such that, in digital circuits, power supply charging or discharging takes place with either the pMOS or nMOS transistor operating in the linear region. In this region MOSFET current flow is given by the following expression

$$I_{ds} = (V_{gs}-V_{th})V_{ds} - V_{ds}^2/2$$

in which there is a clear dependence of drain-source current on drain-source voltage. The equivalent circuit does not make any provision for such dependence; current flow is independent of supply voltage.

In the initial circuit simulation of the processor cell, the resultant current flow data are for an ideal power supply but in assessing the effects of large numbers of simultaneously switching cells, power supply voltage *integrity* will be degraded and the array processor cells will *realign* themselves to a new supply voltage. Supply

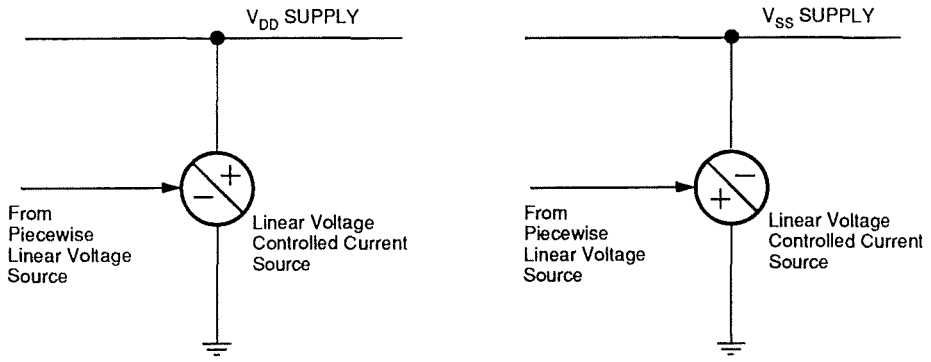


Figure 3.13 Processor Cell Load Equivalent Circuits

voltage most often will be reduced thereby causing a reduction in supply current but provision must also be made for the case where momentarily it may increase thereby causing an increase in current flow.

This *supply feedback mechanism* was modelled with the modified equivalent circuit shown in Figure 3.15. The circuit models the mechanism as follows.

In short, the non-linear voltage-controlled voltage source produces a signal which is related to the difference between the actual supply potential and an ideal reference source. This signal is used to modulate the gate of a MOSFET used in combination with a fixed resistor to divide the amplitude of the piecewise linear source in the original equivalent circuit. In the design of this modified equivalent circuit, it is important to bear the following of its features in mind.

Firstly, it is crucially important to arrange for the MOSFET always to be operating in

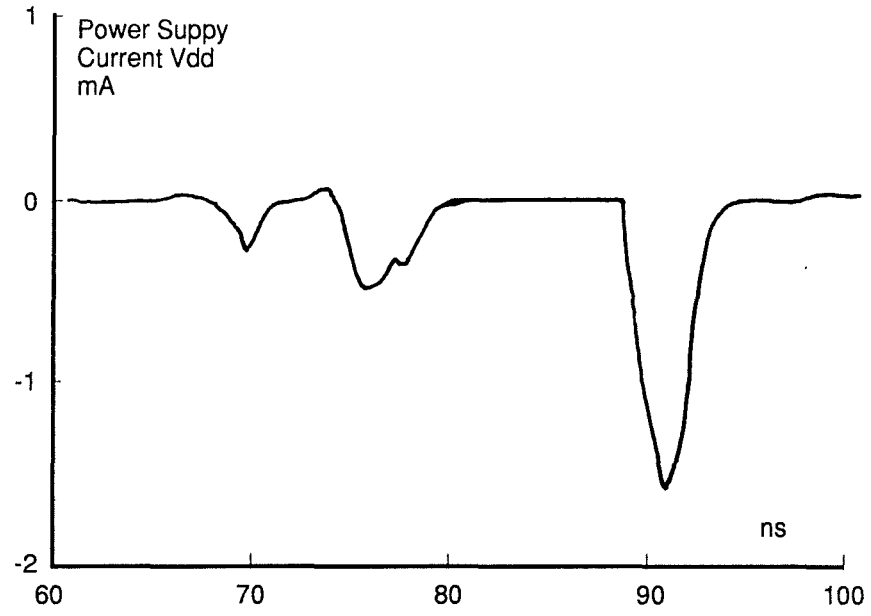


Figure 3.14(a) Processor Cell Power Supply Current (Vdd)

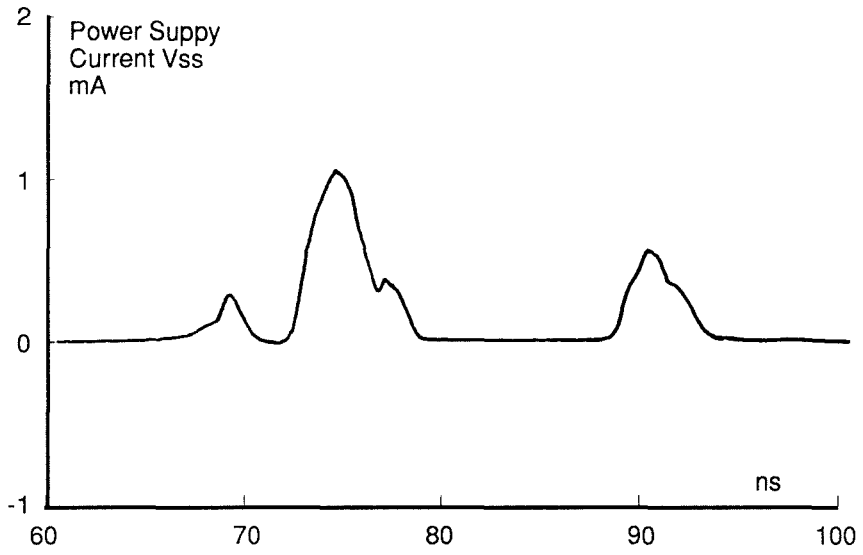


Figure 3.14(b) Processor Cell Power Supply Current (Vss)

the linear region. If this transistor saturates, the piecewise linear voltage source will not undergo a *continuous* reduction in amplitude but rather it will be *clipped* at a level equal to the saturation voltage of the transistor. In order that this effect be avoided, the piecewise linear voltage source was scaled down so as to avoid large drain-source voltages on the MOSFET. The source was scaled down by a factor of ten thereby reducing drain-source voltages to a few hundred millivolts. Clearly, this reduction necessitates an increase in the conductance of the current source.

Secondly, if the transistor has a non-zero threshold voltage, then the piecewise linear voltage source is reduced in voltage by the value of the threshold voltage; in order to avoid this effect the threshold voltage was set to zero volts.

The voltage-controlled voltage source, used as the gate potential for the MOSFET is controlled by the deviation of, for example, the Vdd supply from its ideal value of five volts. The dependence of this gate potential on the deviation of the power supply from its ideal value cannot be expressed linearly but as a sixth-order polynomial in the deviation.

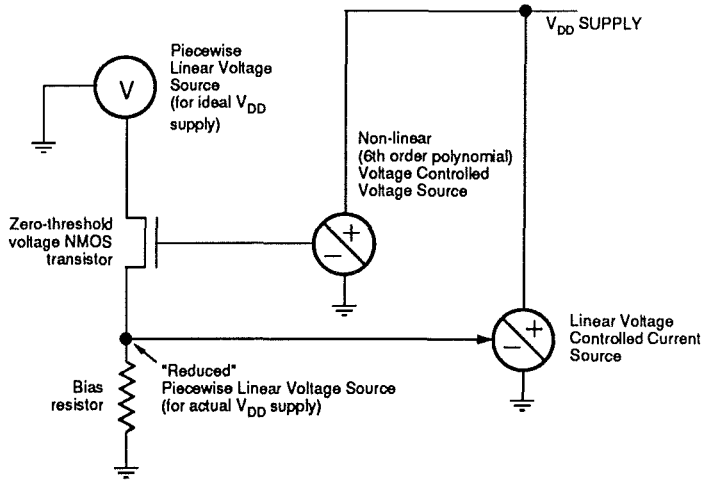


Figure 3.15 Equivalent Circuit with Negative Feedback

Although, by definition, MOSFET transconductance depends linearly on Vds while in the linear region, the rather complex dependence of supply current on supply voltage is manifest in the expression below for current flow in the linear region of a zero-threshold voltage MOSFET. This complex dependence influenced the decision how best to generate a modulating gate voltage source for the modified equivalent circuit.

$$I_{ds} = V_{gs}V_{ds} - (V_{ds}^2/2)$$

$$g_m = (\partial I_{ds}/\partial V_{gs}) \mid V_{ds} \text{ constant}$$

The developed method consisted of SPICE simulating the supply current for a range of fixed supply voltages and noting the reduction in supply current for each of the reduced supply voltages.

The extent to which the supply current was reduced is then the extent to which the linear voltage source as used in the original equivalent circuit needs to be divided. The gate-source voltage, on the zero-threshold MOSFET, was then calculated and checked through simulation.

This resulted in a mapping between deviations in supply voltage and MOSFET gate voltages required to divide the original linear voltage source thereby to effect an appropriate reduction in supply current.

The mapping was expressed as a system of five polynomials; each polynomial corresponding with one of the gate-source voltages necessary to effect the five resistor values for each of the one volt to five volt power supply deviations from ideality. The choice of one volt to five volt deviations from ideality, although convenient is quite arbitrary since this effect clearly is *continuous.* The coefficients of this may be determined by solving a system of five polynomials.

In order to effect the same dependence for the Vss supply, the actual supply was compared with zero volts with the polarities associated with the non-linear voltage-controlled voltage source reversed since, in the case of the Vss supply, the deviation from ideality is in the opposite direction.

The Vdd current flow predicted by the modified equivalent circuit can be compared in Figure 3.16 with the predictions of the original circuit of Figure 3.14(a).

In order to make provision for the case when the positive power supply line should momentarily rise above five volts and/or the negative supply line momentarily fall below zero volts, the circuit was re-calibrated to include ideal supply plus and minus one volt for the positive and negative supplies. A sixth-order polynomial version emerged for each case.

Since the modified equivalent circuit requires three active components the reduction in computing complexity is reduced by 4.66 to 30.

Computational complexity may be kept to an acceptably-low level using this approach and the effect of columns, groups of columns and arrays of simultaneously switching processor cells may be synthesised and assessed.

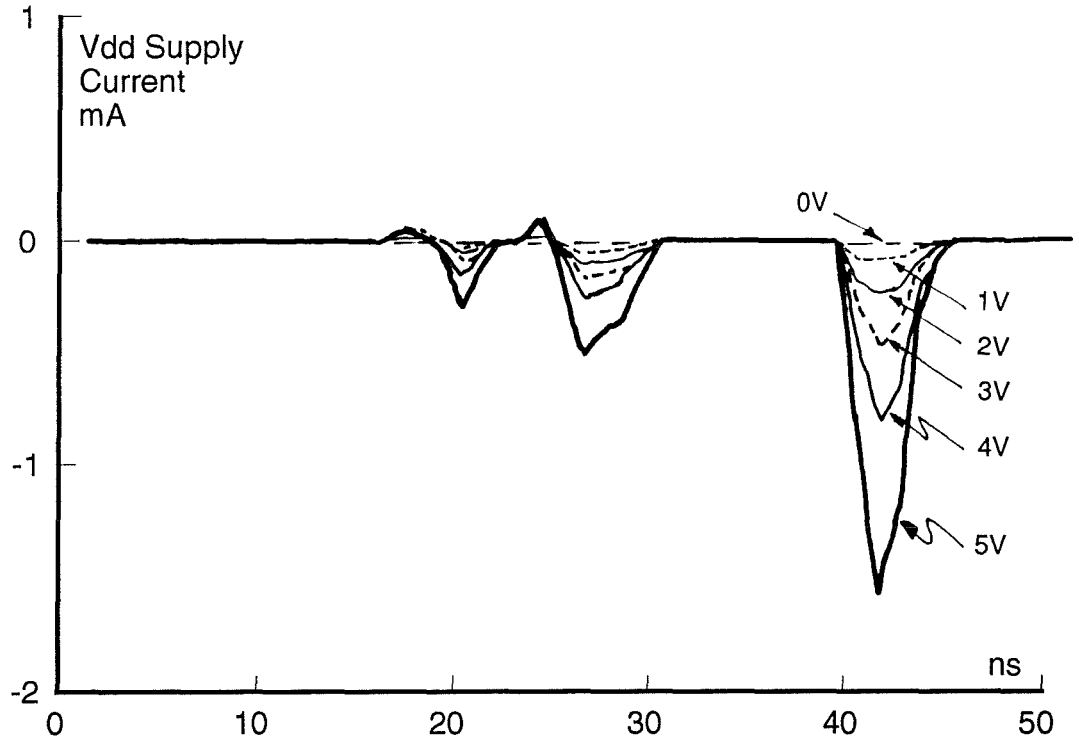


Figure 3.16 Supply Current vs Supply Voltage for modified Equivalent Circuit

3.3.4 Synthesising an Array

In synthesising an equivalent circuit for an entire array of processor cells it is evident, from Figure 3.9, that each of the four columns within any of the twenty groups is subject to equal degrees of clock skew. The array is self-partitioning in this sense allowing independent modelling of each four-column group.

In modelling each four-column group it should be remembered that an alternate-phase clocking scheme is in use so that during any cycle, not all, but half of the processor cells in any column will be active. During a given clock cycle, six out of eleven processors will be active in the first and third columns of each four-column group and five out of eleven processors will be active in the second and fourth columns of each group. During the following cycle, the converse will be true.

In order that this property be featured in a model of the four-column group, the currents drawn by a column of five and of six processor cells was simulated by grouping together five and six of the load-equivalent cells developed earlier. The effects of clock skew within each column was included by using the RCL model used in assessing the performance of the local distribution scheme.

Having simulated the current flow associated with each of a column of six and of five processor cells, it is necessary to determine the skew and supply current transients associated with each column driver. This was done by circuit simulation of a column driver with four column interconnect lines as output load. Power supply current transients for a column driver are shown in Figure 3.17.

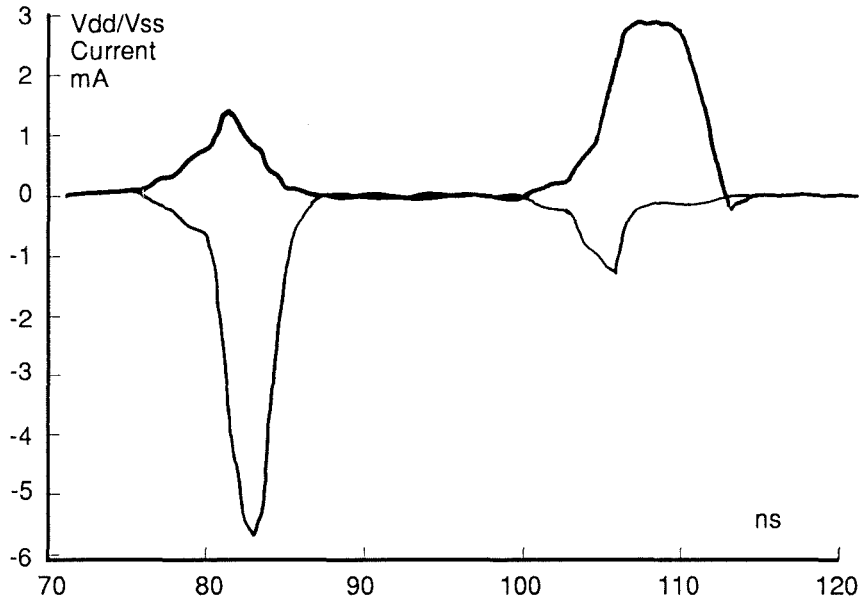


Figure 3.17 Vdd/Vss Current for Column Driver

These were combined with the results for a column of six and of five processor cells, by appropriately delayed linear superposition. The resultant current flow for the four-column group of processor cells is shown in Figure 3.18. This was modelled using a single equivalent circuit in the same way as for the single processor cell so further-reducing the required computing complexity by 117 times. The number of synchronously-active processor cells in a four-column group, 30, raised to the power of 1.4.

In order that these four-column group equivalent circuit could be combined to form an entire array, the effect of clock skew along the edge of the array driver was included by combining twenty such circuits with the horizontal section of the RCL model shown in Figure 3.9.

Power supply current flow for the entire eleven by eighty array of processor cells is shown in Figure 3.19. This was modelled with a single equivalent circuit requiring three active components thereby representing a reduction in required computing complexity of 151,474. The total number of transistors in an array, 30088, raised to the power of 1.4 divided by the number of active components in the equivalent circuit for the Vdd and Vss supply currents, 6, raised to the power of 1.4.

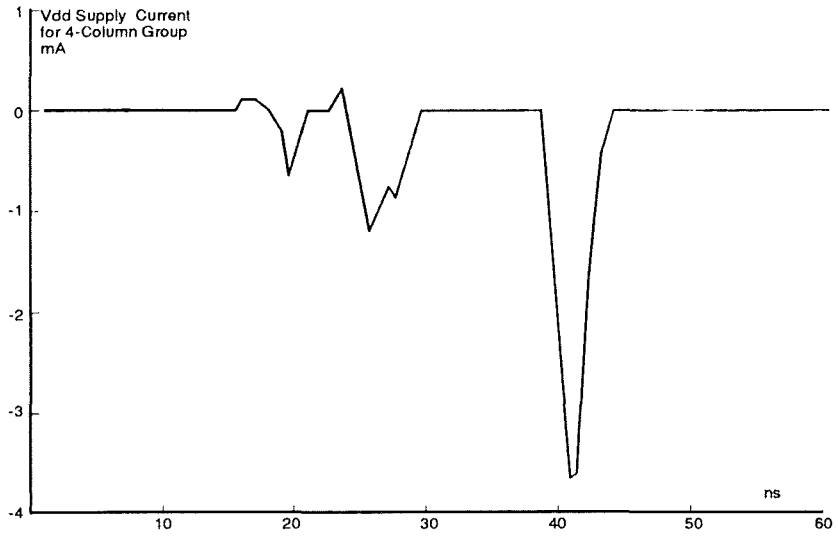


Figure 3.18(a) Vdd Supply Current for Four Column Group

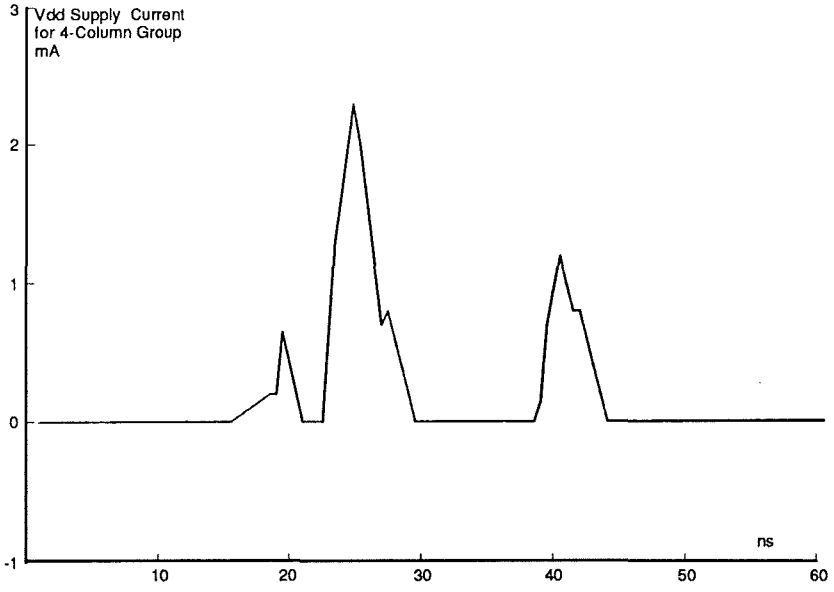


Figure 3.18(b) Vss Supply Current for Four Column Group

3.3.5 The Power Distribution Network

Before equivalent current flow circuits can be used to assess the extent to which power distribution noise is generated, a transmission line model of the power distribution network along with peripheral package-related resistive, capacitative and inductive components needs to be modelled. It is important, in the design of this part of the model, to describe the electrical characteristics of the distribution network in sufficient detail so as to predict accurately the network response. In this case, as in

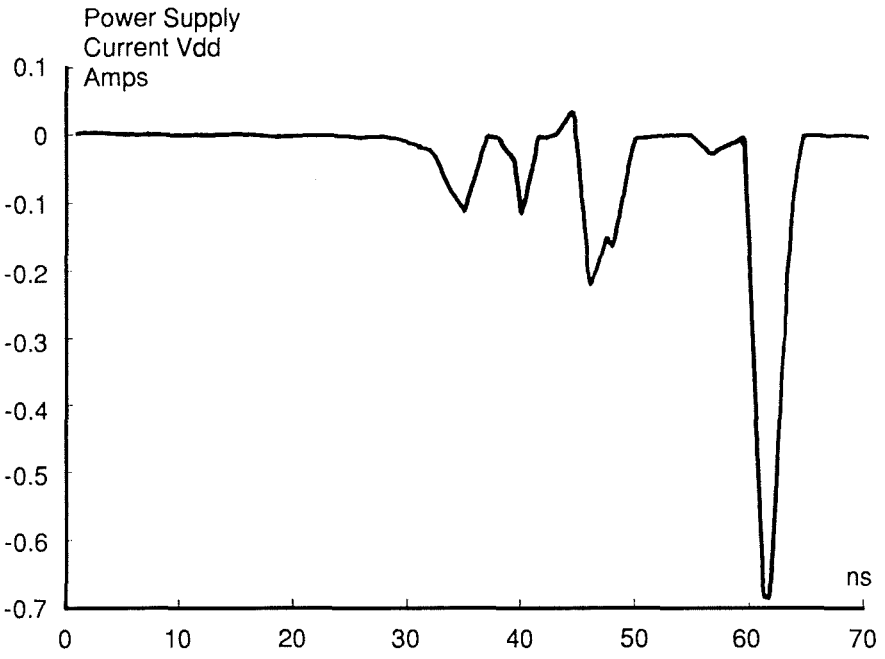


Figure 3.19(a) Vdd Supply Current for Entire Array

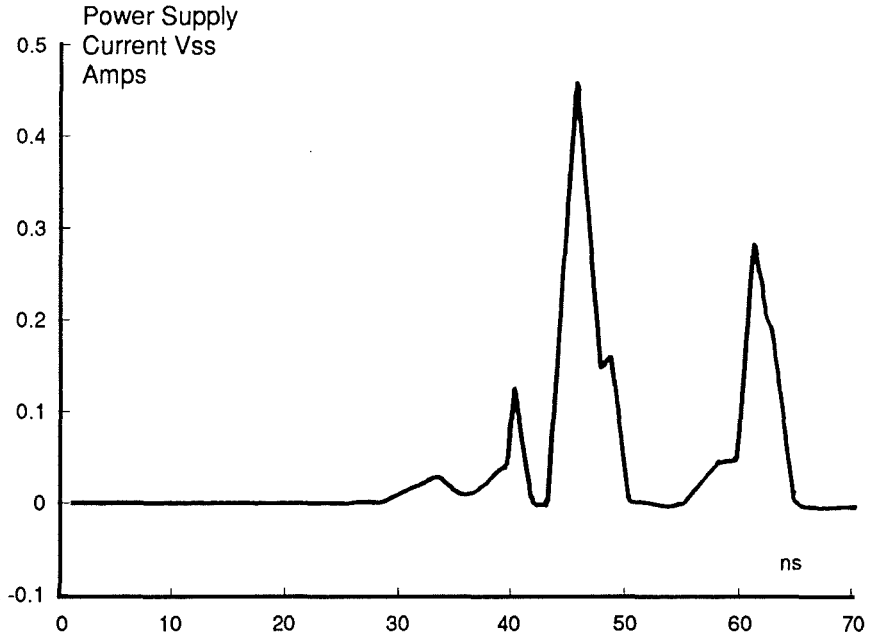


Figure 3.19(b) Vss Supply Current for Entire Array

the case of the global clock distribution analysis, each *section of interest* is assumed to be one array of processors so that an RCL ladder element for each 0.5 millimetre of interconnect.

The chosen power distribution network topology is shown in Figure 3.20. Remembering that the intended use of this simulation model is to assess the extent to

which power distribution *may limit* the performance of the circuit, it is clear that it should at least be assumed that an entire interconnect layer is devoted to power distribution.

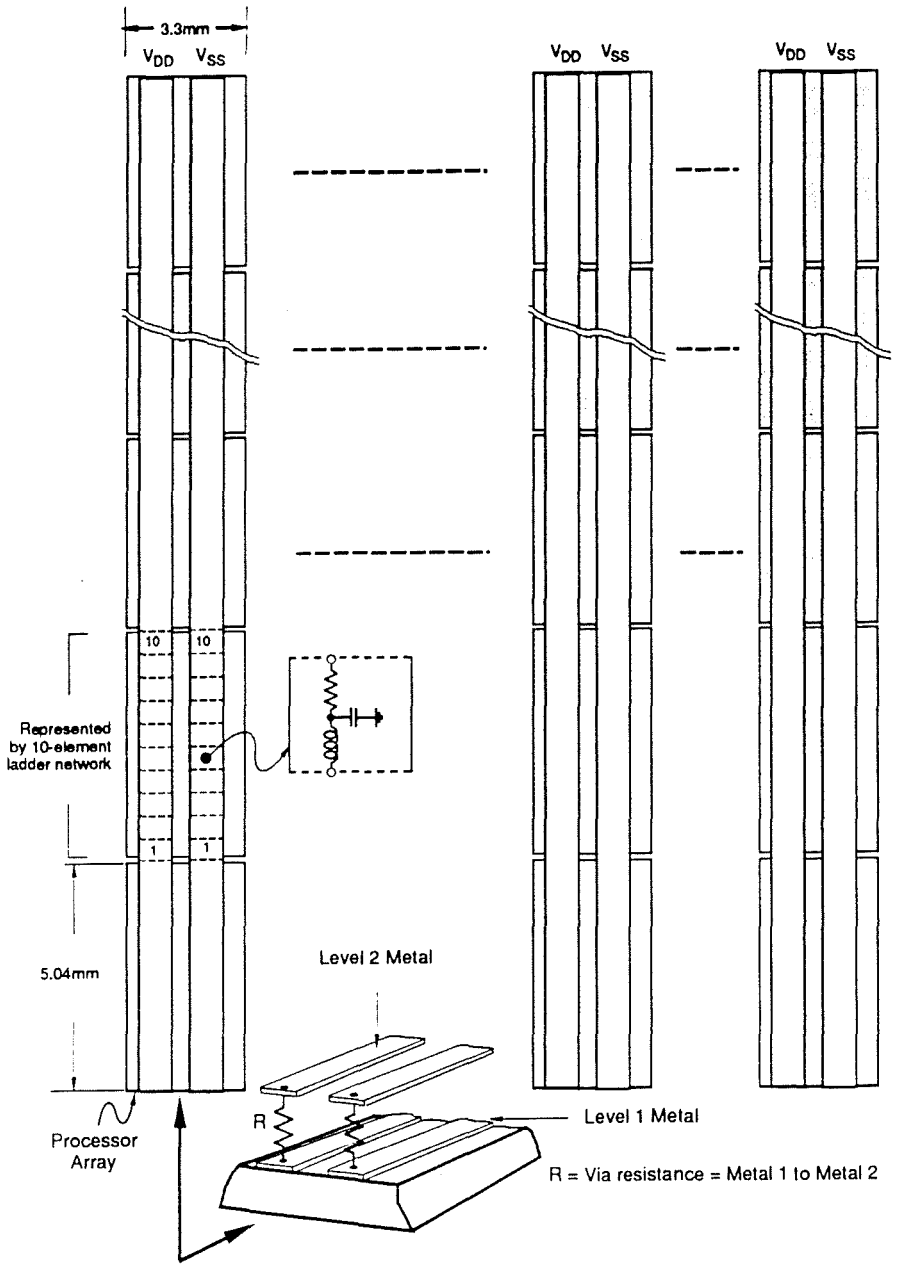


Figure 3.20 Power Distribution Network Topology

Since each array is around 3.3mm high, it is assumed that each of the Vdd and Vss distribution lines is around 1.5mm wide in order that track resistance be minimised and that track capacitance be maximised.

Further, it must be borne in mind that it is current practice to distribute global

signals on a second layer of metal which is connected, at regular intervals, through *interconnect vias*, to the first layer. In short, the power distribution scheme has become a two-level-metal distribution network as shown. Notice that the first and the second layers have been chosen to have the same topologies and that vias will be introduced at intervals of 100 microns.

The first assumption is justifiable since, in the likely event of the first layer having a topology that is made up of many narrow and *randomly* distributed tracks, their composite electrical characteristics, at each half millimetre interval, will differ from that of the second layer only in the relative increase in peripheral capacitative components. These peripheral components are believed not to be significant at two-micron line widths. Their significance is shown in the sensitivity analysis of section 3.5.2. The second assumption is broadly in line with current practices.

Package-related electrical components include the resistance, capacitance and inductance of the bond wire from silicon to package, the package pad, on to which the bond wire is connected, and the package pin. No values could be assigned to two of these nine parameters; these were package pad resistance and inductance. They are thought to be negligibly small.

Values for the remaining six parameters were chosen so as to be comparable with those values listed in Figure 1.14. The chosen values are:

Rpack	=	0.2 ohms
Cpack	=	2pF
Lpack	=	7nH
Rbond	=	0.01 ohms
Cpad	=	2pF
Cbond	=	1pF
Lbond	=	1.25nH

These components were chosen to represent electrical parasitics associated with modern packaging technology and were configured as shown in Figure 3.21.

With all of the model components defined, the equivalent circuits for each of Vdd and Vss supplies were connected to the distribution network. So that the load would be distributed evenly, each equivalent circuit was partitioned into five equal parts and connected to every second ladder element in the ten-element section representing the power distribution network for each array.

Since it is neither the instantaneous value of Vdd(t), nor of Vss(t), which affects circuit performance, but the instantaneous *difference* in these signals, a method of deriving this instantaneous difference signal was created in the form of the voltage-controlled voltage source configured as shown in Figure 3.22.

Since these model additions introduce five times as many active components plus one voltage-controlled voltage source, computing complexity is increased 31/6 raised to the power of 1.4, 9.97, times so that overall the reduction in computing complexity falls from 151,474 to 15,200.

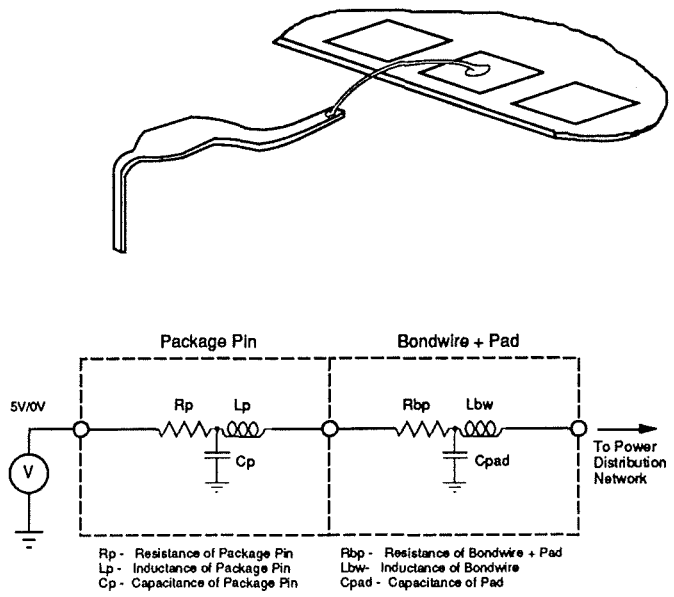


Figure 3.21 Package-related Parasitics

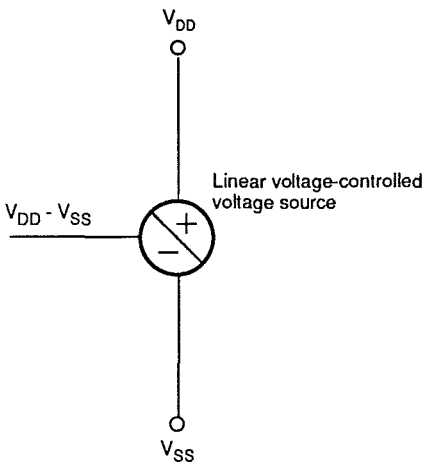


Figure 3.22 Voltage Subtractor

3.3.6 Noise Predictions

With the simulation model complete, the remaining objectives are to use the model to assess the magnitude, and predict the nature of, power distribution noise and the extent to which it places a constraint on circuit performance.

In pursuing this objective, the combined effect of up to seventeen single-array elements was simulated. Beyond this level, there emerged simulation problems associated with the amount of data generated internally by SPICE. Although these problems could have been overcome by, for example, reducing the degree to which the equivalent circuits were distributed evenly over the supply network, the results obtained revealed trends that lent themselves to confident extrapolation.

In order that a clear perception of the predicted trends in power distribution noise for

the seventeen cases spanning from that associated with a single array through to that associated with seventeen arrays, a sample range of the results for one, five, eight, thirteen and seventeen arrays is shown in Figures 3.23, 3.24, 3.25, 3.26 and 3.27. These arrays correspond to integration levels of 30,000, 150,000, 240,000, 390,000 and 510,000 "synchronously-active" devices.

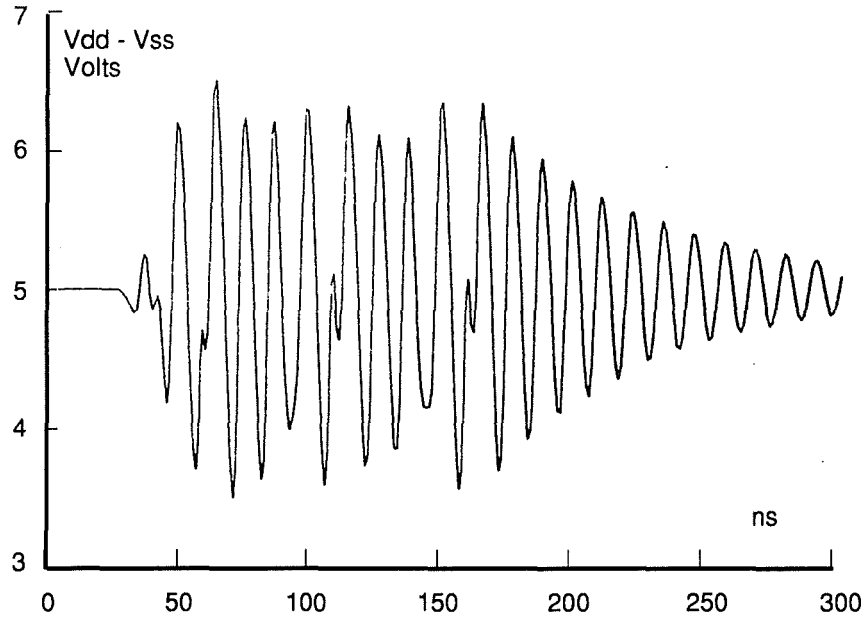


Figure 3.23 Instantaneous Difference in Vdd and Vss Single Array

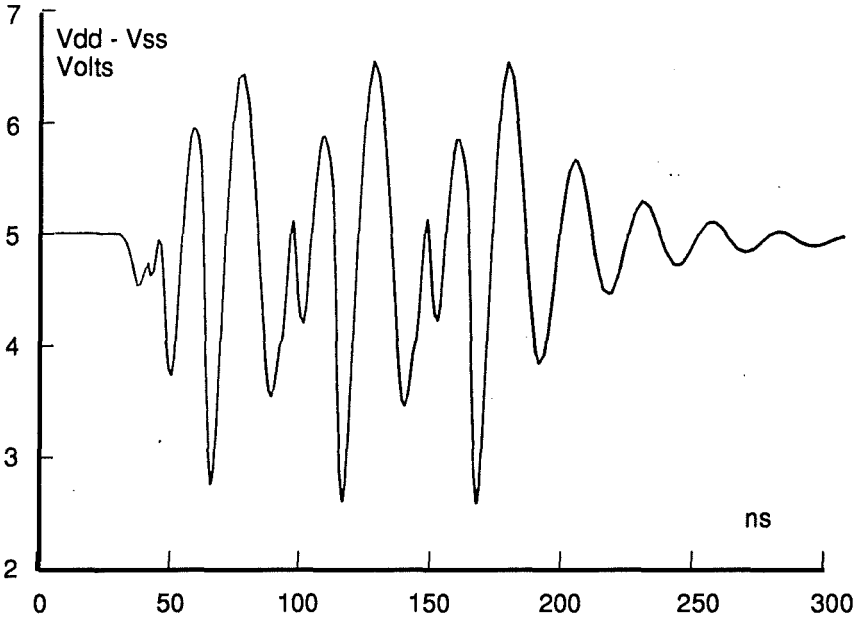


Figure 3.24 Instantaneous Difference in Vdd and Vss Five Arrays

The figures show the predicted "instantaneous difference" in the positive (Vdd) and negative (Vss) power supplies over an active period of three clock cycles (150ns) and for the physical mid-point of each of the Vdd and Vss supply lines.

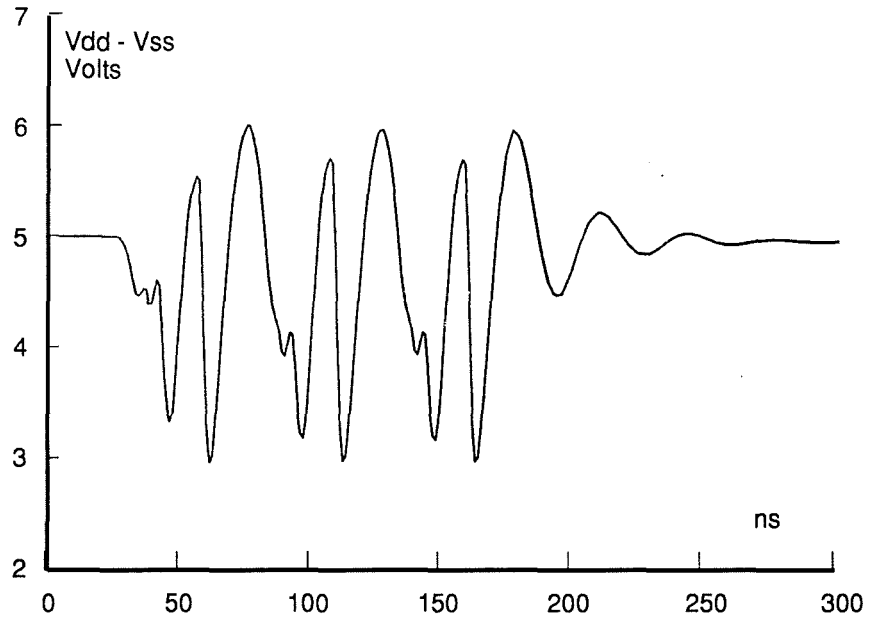


Figure 3.25 Instantaneous Difference in Vdd and Vss Eight Arrays

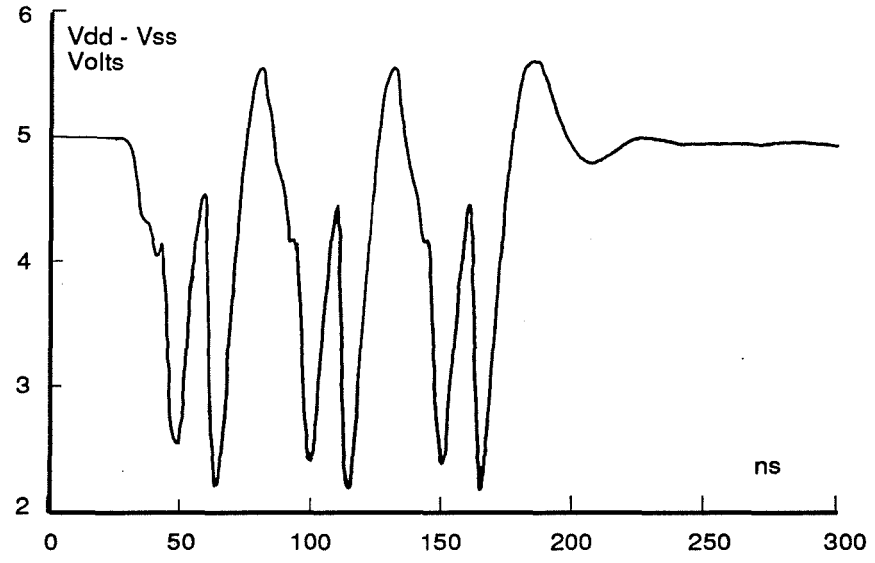


Figure 3.26 Instantaneous Difference in Vdd and Vss Thirteen Arrays

The exponentially decaying sine waves, most obvious in the single-array case, reveal that the distribution networks are *second-order linear* networks whose transient response is, in general, given by [d09].

$$V(t) = \exp(st) = \exp(\alpha t).\exp(j\omega t), \text{ where } \alpha < 0$$

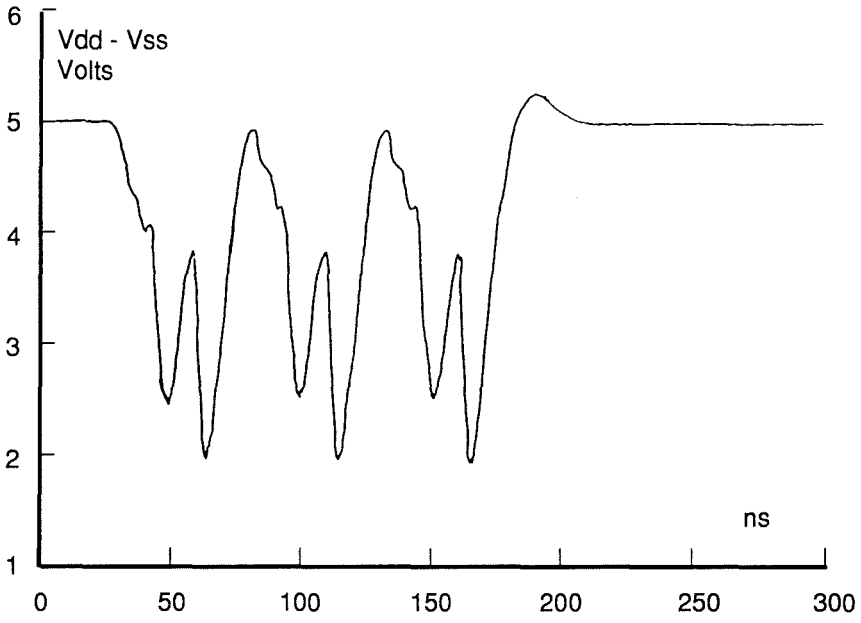


Figure 3.27 Instantaneous Difference in Vdd and Vss Seventeen Arrays

In this case, α has been shown in section 1.2.3 to be equal to $(-Rs/2Z_0)$ and is observed gradually to increase as distribution networks become larger. As the lossy integrated circuit metallisation interconnect becomes longer the source resistance of the equivalent transmission line increases while its characteristic impedance, $Zo=\sqrt{(L/C)}$, remains constant.

The instantaneous difference in the supply voltages, hereinafter referred to as *voltage integrity*, has the form described by a second-order differential equation in current with respect to time with L, C and R as equation coefficients [d10]. This form is confirmed in Section 3.5.

$$dV/dt = Ld^2i/dt^2 + i/C + Rdi/dt$$

The model has predicted that with smaller *LSI-sized* circuits the exponential attenuation period may exceed the clock period (50ns) and consequently result in transient supply voltages which are not predicted by the above equation. This behaviour consequently is dependent on clock frequency and circuit size and therefore may be thought of as a *quasi-resonant* phenomenon.

In order better to illustrate the underlying trends in Figures 3.23 to 3.27, minimum and maximum voltage integrity are graphed against number of arrays in Figure 3.28. It is then clear that, with the exception of those circuits whose exponential attenuation is comparable to the 50ns clock period, the results for both minimum and maximum voltage integrity follow a near-linear curve thereby exhibiting compatibility with the above equation in L, C and R.

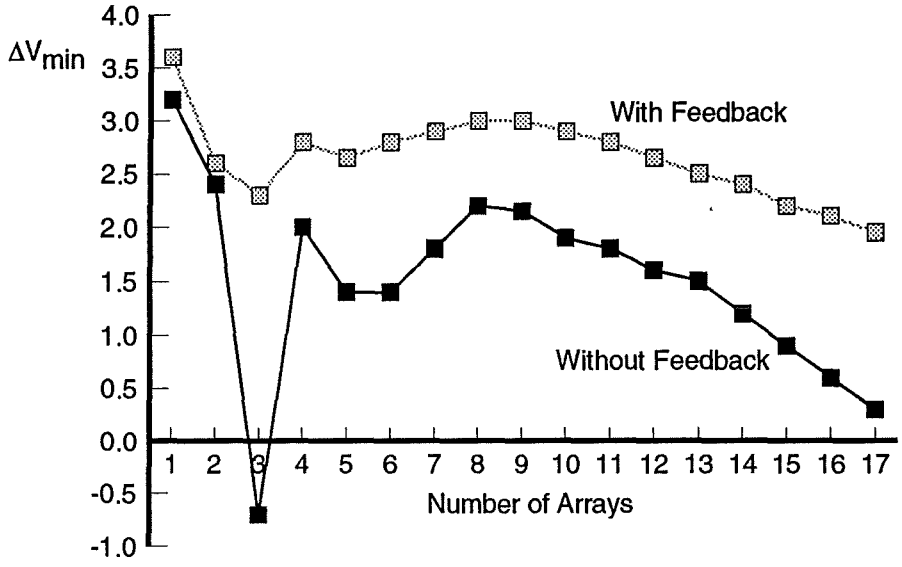


Figure 3.28(a) Minimum Voltage Integrity vs Circuit Size

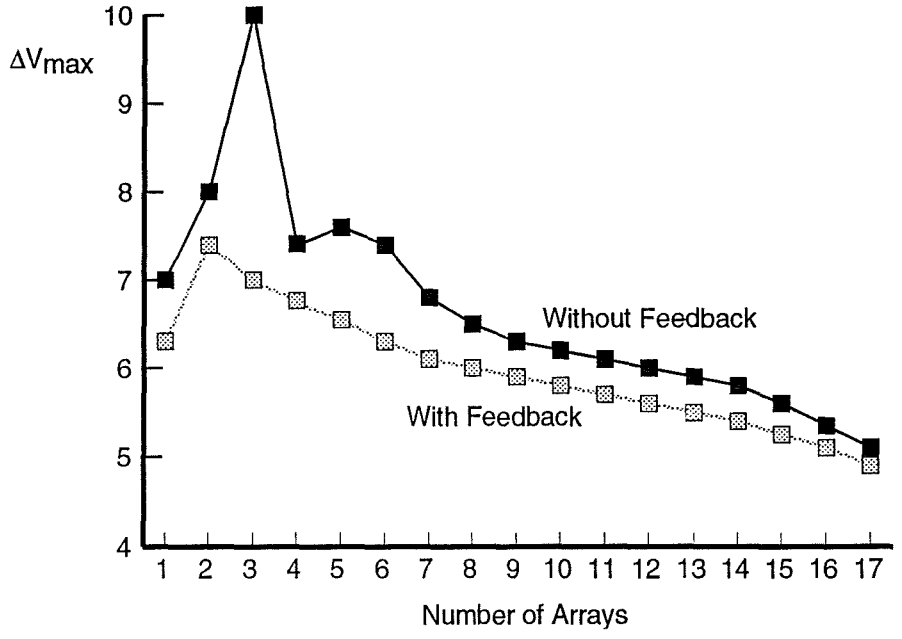


Figure 3.28(b) Maximum Voltage Integrity vs Circuit Size

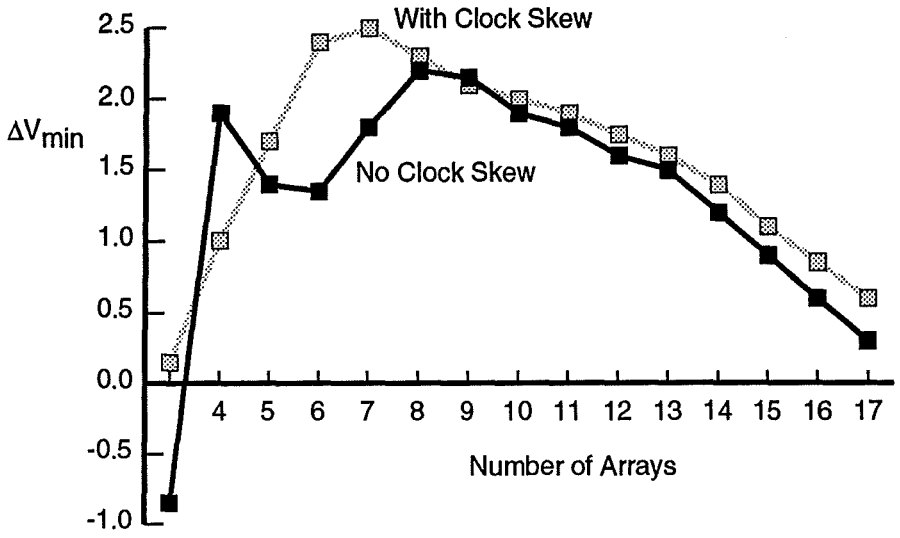


Fig 3.29(a) Global Clock Skew - ΔVmin

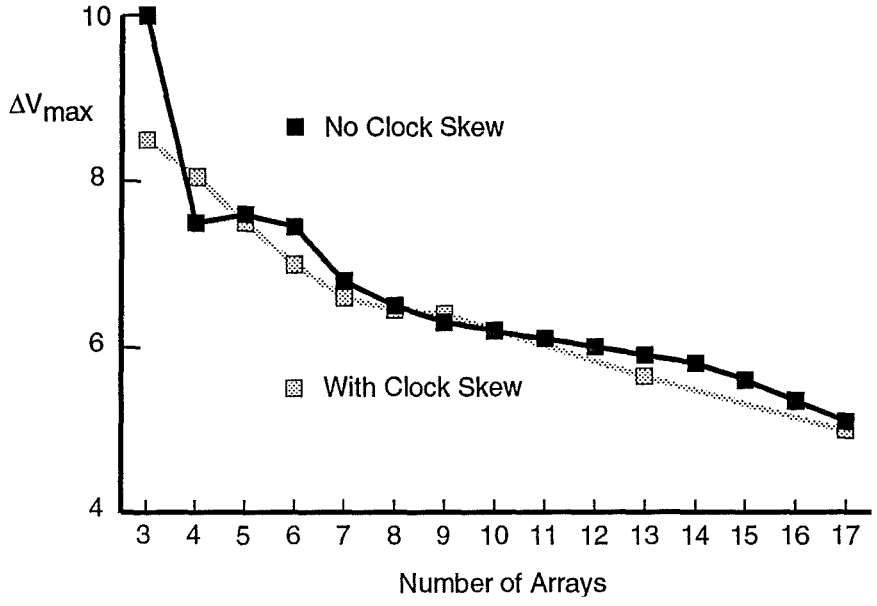


Fig 3.29(b) Global Clock Skew - ΔVmax

Notable, also, is the extent to which the dependence of supply current on supply voltage, embraced by the supply feedback mechanism described in section 3.3.3, affects voltage integrity. In the case of minimum voltage integrity, an *improvement* approaching 100% of their value without supply feedback is observed for the largest circuit with *improvements* for the other circuits approximately in proportion to their relative size. The effect is much less marked in the case of maximum differences since there is much less scope built into the simulation model for *equivalent circuit realignment* and, as can be seen from the upper curve in Figure 3.28(b), much less need for such realignment.

A subsidiary objective of this exercise is to determine the extent to which clock skew affects voltage integrity levels. As discussed earlier, clock skew exists at two levels within this architecture: it exists at the level of a single array of processors, previously referred to as *local* clock skew; and it is introduced, in a form previously referred to as *global* clock skew, as the clock is broadcast along the linear array of processor arrays.

Local clock skew has been found to have no resolvable effect on voltage integrity levels.

In the case of global clock skew as described in section 3.3.2, it is evident from Figure 3.29, that associated reductions in i, di/dt and d^2i/dt^2 do have a resolvable effect.

It can be seen that the observed *quasi-resonant* behaviour, associated with circuits of less than eight arrays, has been reduced and for the larger arrays, it is clear that a gradually more optimistic prediction for minimum voltage integrity emerges. For reasons of computational complexity, the results of Figure 3.29 do not include the supply feedback mechanism of section 3.3.3.

3.3.7 Non-standard Technology

The analysis, so far developed, has considered power distribution using standard two-level aluminium metallisation with peripheral bond wire connections to the outside world; so called *standard technology.* It may prove instructive if a similar analysis, based on the electrical characteristics associated with non-standard, but emerging, power distribution technology, were undertaken. Examples of non-standard power distribution technology include copper electroplating and non-peripheral bonding.

Electroplated copper conductors have been realised by Barrett [d11]. The emergent process sequence allows the deposition of a layer of copper metal on the passivation layer of an, otherwise standard, integrated circuit. The process allows deposition of copper of up to twenty microns thickness with sheet resistivities at around one fortieth of standard aluminium metallisation. In addition to these substantial reductions in resistivity, there are increases in capacitance and decreases in inductance, each due to the substantial increases in metal thickness relative to those normally found with standard aluminium metallisation. In the analysis that follows, it is assumed that copper is electroplated on to the upper passivation layer and that it is connected to the second layer aluminium metallisation which, as described previously, is connected in turn to the first layer aluminium metallisation.

Referred to in section 1.4.6, non-peripheral bonding is made possible through pin grid *flip-chip* mounting technology. This technique is facilitated if the circuit in question is fault-tolerant because then the non-peripheral power supply site may be treated effectively as a large processing defect and therefore be circumvented.

In addition and in contrast to each of these radical process developments, it is possible, during the course of standard packaging processes, simply to include one additional bond wire for each of the positive and negative power supply bond pads. As discussed previously, this would have the effect of halving bond wire resistance and

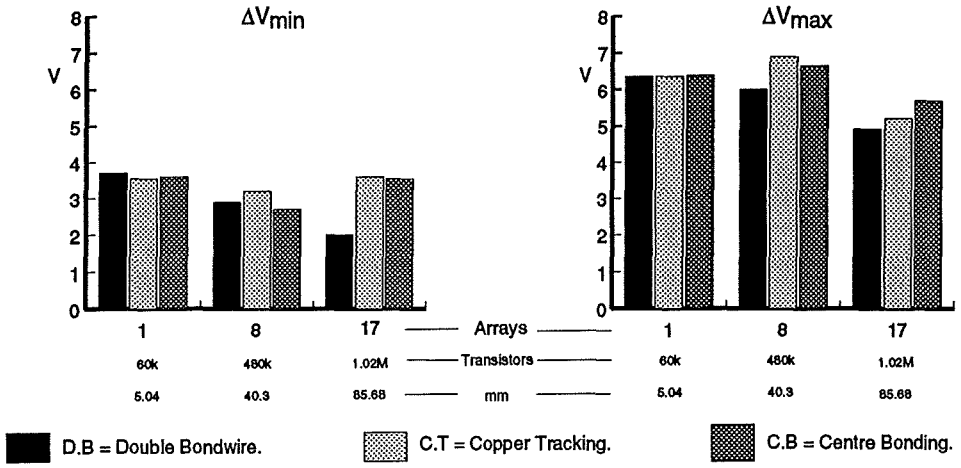


Fig 3.30 Maximum & Minimum Voltage Integrity for Non-standard Technologies

inductance while doubling the effective bond wire capacitance; all three effects are expected to assist in the provision of reduced power distribution noise.

Noise analyses were repeated for each of these non-standard power distribution technology scenarios. The results obtained for three micron electroplated copper distribution tracks, bonding to the centre of the integrated circuit and double-bonding at the integrated circuit periphery, as applied to the seventeen-array circuit operating at 30MHz, are shown in Figures 3.30. These results will be analysed quantitatively in section 3.4.2.

The seventeen-array case, operating at 30MHz, was chosen since it has highest dissipation. Results for this case will best represent the *potential* improvements in voltage integrity that may be achieved through the use of these non-standard power distribution technologies. The simulation model was adapted for 30MHz operation by inverse-proportionately scaling the time separating each of current peaks. In order to effect an upwards shift in the operating frequency from 20MHz to 30MHz, the time separating the current peaks for each of the positive and negative supply lines was scaled by a factor of 0.67.

3.4 Performance Implications

3.4.1 Assessment Methodology

Further interpretation of the above predictions for power distribution noise is difficult. It is unjustified, for example, to assert that a degradation in voltage integrity to a level below any particular level, will result in unreliable circuit operation or, in the extreme, a non-functional circuit. Circuit sensitivity to voltage integrity is dependent on circuit type and design methodology. Undertaking a transient circuit simulation of the performance of the chosen processor cell in such a noisy environment, although feasible, is inappropriate since this would lead to results that are relevant specifically

for the processor cell design under investigation. What is required is to develop a performance assessment methodology that is applicable to *range* of digital CMOS integrated circuits.

In order to develop a method of deriving more general implications, a more general view of what has been undertaken is adopted.

A simulation model has been developed that can be used to predict the effect of a large number of digital circuits, operating simultaneously, on the positive and negative power supply lines. A value for the difference in the levels of the positive and negative power supply lines, that is transient in time, has emerged which has provided a measure of the power supply degradation. What is the reason for this predicted degradation? It is a direct consequence of the electrical characteristics of the power distribution network with its associated resistance, capacitance and inductance. In short, it is a consequence of the process and power distribution technology.

A parameter that is widely used as a general figure of merit for a given process technology is *gate-delay*. The problem, therefore, can be reduced to that of developing a methodology for assessing the extent to which gate-delay is degraded by the predicted levels of voltage integrity obtained from the simulation model. In developing such a methodology a gate must be chosen that is representative of a wide range of digital circuits.

The obvious choice is the CMOS inverter since it is generally representative, at an elemental level, of many CMOS digital circuits and with reference to Figure 3.31, it is clear that the systolic correlator processor cell has an abundance of simple inverter structures thereby making this choice of gate particularly relevant. The problem of developing a technique for gate-delay assessment was addressed next.

3.4.2 Performance Assessment

The predicted value for the instantaneous difference in the power supply rails, transient in time, was separated into its constituent positive (Vdd) and negative (Vss) components so that each could be applied to the positive and negative terminals of the inverter thereby subjecting the inverter to the levels of power distribution noise predicted for each of the increasingly larger systolic arrays in order that the first inverter be appropriately loaded.

Gate-delay degradation associated with two serially connected inverters was assessed and results for the case of seventeen arrays operating at 20MHz and at 30MHz are shown in Figures 3.32 and 3.33. Since the circuitry begins to dissipate on clock rising edges and finishes on clock falling edges, the relevant parameter for the assessment of performance degradation is fall-time. This is a consequence of an inherent assumption of positive logic in the degradation assessment methodology.

Voltage integrity predictions, obtained for non-standard power distribution technology variants were used similarly to assess their effect on gate-delay performance. Results

for the relative degradation in fall-times, associated with each technology variant, are shown graphically in Figure 3.34.

It is clear that, for the seventeen-array case operating at 30MHz, improvements in fall-time degradation of around 85% are predicted for each of electroplated copper tracking and for centre-bonding while double-bondwire connections offer improvements in fall-time degradation of less than 10%.

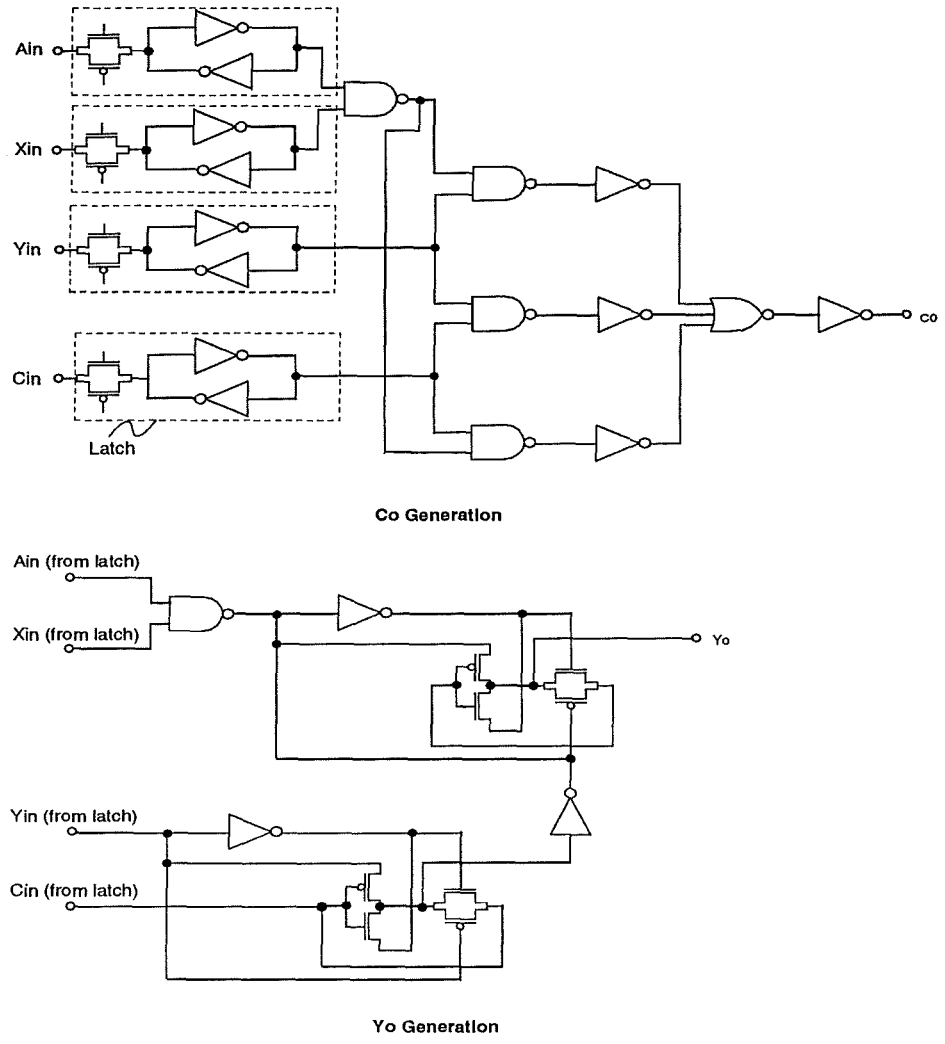


Fig 3.31 Processor Cell Inverter Structures

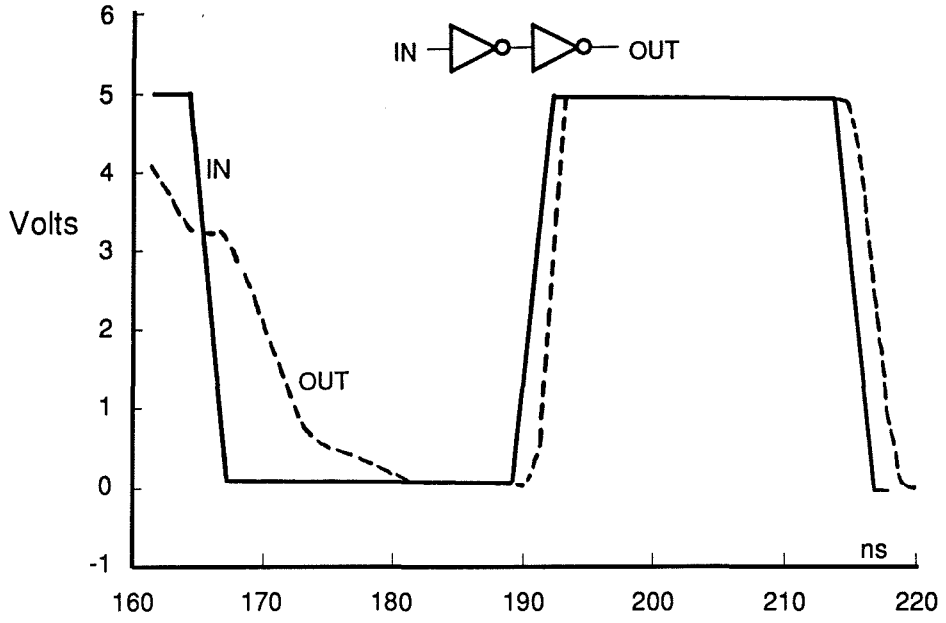


Fig 3.32 Gate Delay Degradation for Seventeen Arrays (20MHz)

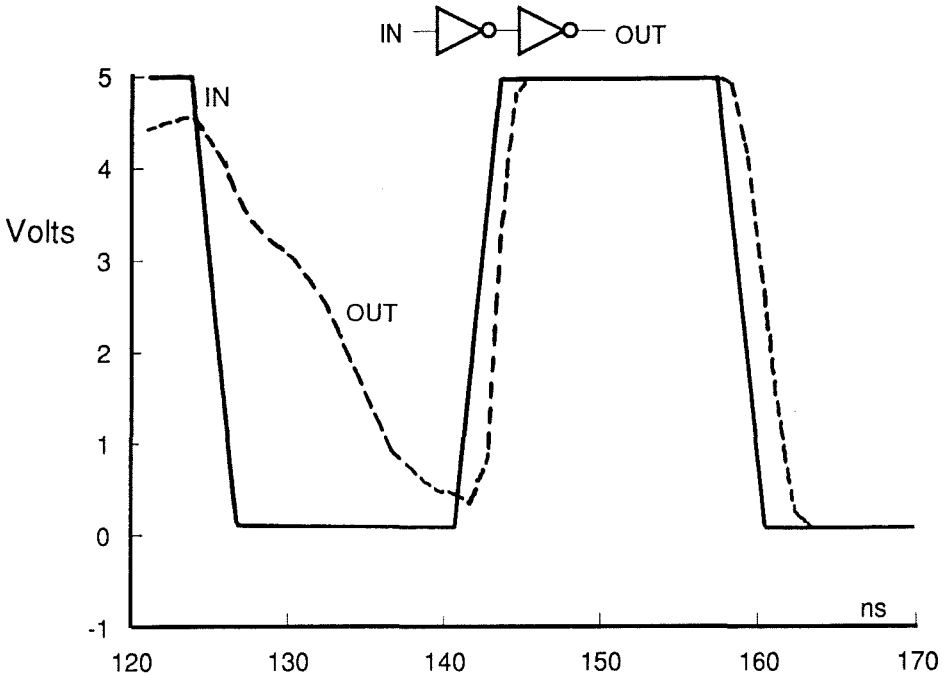


Fig 3.33 Gate Delay Degradation for Seventeen Arrays (30MHz)

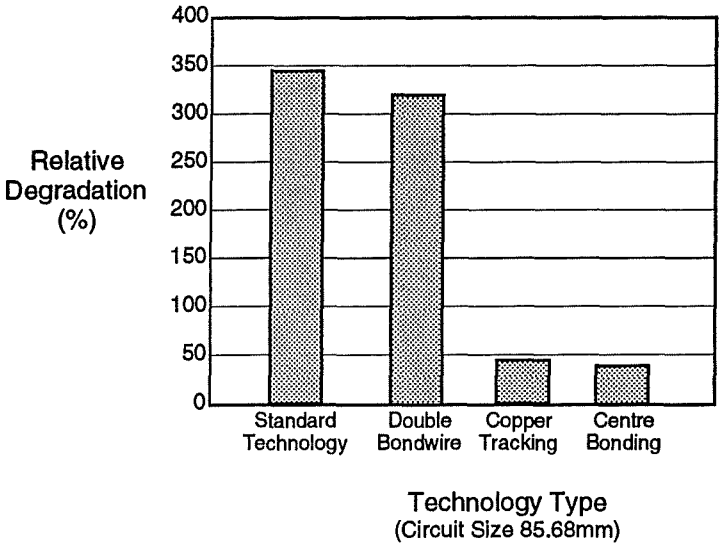


Fig 3.34 Relative Fall Time Degradation for Seventeen Arrays operating at 30MHz

3.5 Noise Model Sensitivity

3.5.1 Assumption-related Issues

In developing the above simulation model, many assumptions and design choices have been necessary. In order that the emergent results have a consequential meaning for present processing technologies, for example, many assumptions and values have been included that relate to a specific process technology. The purpose of this section is to assess the extent to which the results that have emerged are dependent on these assumptions and model design choices.

Working backwards through the model structure, the following is a list of features that have a direct bearing on the predictions for overall voltage integrity emerging from the simulation model:

1) Package-related resistance, capacitance and inductance.

2) Distribution network resistance, capacitance and inductance.

3) First-layer distribution network resistance, capacitance and inductance (independently of second-layer).

4) Distribution network metallisation inter-layer via resistance.

5) Load as observed by the network.

6) Test vector transition times.

The above are concerned with parametric values which have been included in the simulation model. Other assumptions and design choices that have been necessary, but which are less directly part of the model, will be discussed separately.

In the first case, listed above, the parameters involved are: bond wire resistance, capacitance and inductance; package pad capacitance; and package pin resistance, capacitance and inductance. Each of these values independently were varied by -75% through to +100%, in 25% increments, and the resultant effect on the predicted voltage integrity is shown graphically in Figures 3.35, 3.36, 3.37, 3.38, 3.39, 3.40 and 3.41. This variation was not undertaken for all seventeen circuit sizes but, selectively, for each of the single-array, the eight-array and the seventeen-array cases thereby to ascertain any inherent trend.

The results indicate, clearly, that the predicted voltage integrity levels for each of the single-array, the eight-array and the seventeen-array circuits do not depend, to any resolvable extent, on bond wire resistance, bond wire capacitance, on package pad capacitance or on package pin capacitance.

Dependence on bond wire inductance, package pin resistance and inductance is predicted. Figure 3.41 illustrates the underlying trend for each of these dependencies; namely that they are strongest for the single-array case, weaker for the eight-array and weakest for the seventeen-array case. This result is consistent with the earlier prediction that, as circuits become larger, the quasi-resonant nature of the power distribution network becomes less prevalent.

The second case, listed above, involves the independent variation of the resistance, capacitance and inductance of both the first and second layer metallisation network by -75% through to +100% in 25% increments. The results for each of the single-array, the eight-array and the seventeen-array cases are shown in Figures 3.42, 3.43 and 3.44.

The results for variations in metal resistivity, shown in Figure 3.42, indicate a near-linear dependence of voltage integrity for the eight-array and the seventeen-array cases.

The results for variations in network capacitance, shown in Fig 3.43, indicate a non-linear response for the single array case and a second order linear response for each of the eight array and seventeen array cases. The non-linear dependence associated with the single array case is due to the fact that the exponential decay period is comparable with the circuit clock period thereby resulting in a *quasi-resonant* response.

The results of Figure 3.44 indicate no voltage integrity dependence on metallisation inductance thereby implying that the distribution network response is second order due to package related parasitics only. The internal recirculating currents, as described in section 1.5.1, are insufficient to cause any supply perturbation

The third case, listed above, involves the independent variation by -75% through to 100%, in 25% increments, of the resistance, capacitance and inductance of the first layer metallisation network independently of the second layer. The results for each of the single-array, the eight-array and the seventeen array cases are shown in Figures 3.45, 3.46 and 3.47.

The purpose of determining the extent to which the voltage integrity results are sensitive to independent variation of the electrical characteristics of first layer metallisation, with respect to the second, is needed because of the earlier assumption made regarding the relative size and topology of these layers. The nature of the dependence is the same as the previous case in which the electrical characteristics for both layers were varied together. The point to note, for this case specifically, is the absolute value of the dependence.

Results for the case of the eight-array and the seventeen-array cases are tabulated in Figure 3.48. Careful study of these results reveals that independent variation of first-layer metal resistivity reveals that two thirds of the voltage change is associated with this layer and one third with the second. The observed sensitivity is explained by the fact that first layer metal resistivity is twice that of second layer. In the case of metallisation capacitance first-layer metallisation capacitance is half that of second-layer and the trends are reversed.

The fourth case, listed above, simply involves varying the first layer metallisation to second layer metallisation via contact resistance by -75% through to 100% in 25% increments. The results for each of the single-array, the eight-array and the seventeen-array cases are shown in Figure 3.49. It is clear that the simulation model is not sensitive to this parameter.

The fifth case, listed above, involves a variation of the voltage-controlled current source, in each of the equivalent circuits, by -75% through to 100% in 25% increments. The results for each of the single-array, the eight-array and the seventeen-array cases are shown in Figure 3.50.

Clearly, this analysis is aimed at assessing the extent to which the predicted results, depend on the *load* that is placed on the power distribution network. The load may change because of an architectural modification since, as explained in section 3.1, systolic arrays may have varying degrees of *efficiency*. In this analysis, it has been assumed that only half of the constituent processor cells are active at any given time. McCanny and McWhirter [d2] describe a systolic array for correlation in which all of the processors become active synchronously. The load may change also if the process technology that is chosen for circuit implementation is upgraded. The effects of a radical technology change from say, bulk CMOS to SOI-CMOS or to bipolar, cannot be assessed by a such a simple change to the equivalent circuit since, in such cases, the entire current profile would doubtless be changed substantially.

The last case, listed above, involves changing the rise-time and fall time of these test vectors. The results for each of the Vdd and Vss current pulses are shown in Figure 3.51.

The importance of understanding sensitivity to test vector transition times is associated with a missing feedback path in the simulation model. In its present form, it fails to account for the relationship between the predicted increases in gate fall times and the input test vector fall times.

The results indicate that, as fall time is increased, pulse duration is proportionately increased and pulse amplitude decreased. The trend is to be expected since the total

charge required to raise the voltage on a fixed output load is fixed (Q=CV).

In addition to the above features which relate directly to the simulation model, the results that it has produced depend also on those transistor parameters such as transistor threshold voltage, gate oxide capacitance and junction temperature. The values chosen for these parameters in these analyses are representative of modern integrated circuit fabrication processes and operating conditions and the sensitivity of current profiles to each of these parameters is well established.

3.6 Conclusions

3.6.1 Simulation Model

In addressing objective (2) outlined in section 3, a simulation model has been developed to assess the extent to which power distribution noise may limit the integration levels and performance of systolic array integrated circuits of ever increasing size.

The model adopts an equivalent circuit approach to noise modelling that has allowed a reduction in circuit-simulation computational complexity by a factor in excess of 15,000. It has been developed to assess the nature, the magnitude and effect of power distribution noise associated with systolic array integrated circuits that involve the integration of up to 1.020,000 transistors configured as a *linear* array of *LSI-sized* arrays of some 60,000 transistors.

If this linear array is repeated until the array is physically square, then these results are representative of a systolic array occupying 64 sq. cm and involving some twenty nine million transistors.

The model has been used to assess the effect of predicted noise levels on circuit performance for standard power distribution technology common to CMOS integrated circuits. In addition, the potential benefits associated with non-standard but emerging, power distribution technology have been assessed.

The distribution networks have been shown to have a second order linear response. The predicted noise is, except for smaller *LSI-sized* circuits, described by a second order differential equation in LCR.

$$dV/dt = Ld^2i/dt^2 + i/C + Rdi/dt$$

No transmission line effects were predicted.

The model has predicted that, with smaller LSI-sized circuits, exponential attenuation ($\exp(-Rs/2Z_0)$) times may be comparable with the clock period and consequently result in transient supply voltages which are not predicted by the above equation.

Since this effect is dependent on circuit size and clock frequency, it may be regarded as a *quasi-resonant* effect.

Further, the model has predicted that, for the circuits examined, the effects of metallisation are negligible. All exponential responses are associated with package-related inductive elements.

The effects of clock skew are predicted overall to be very small and insignificant at loads of below 165 Ohm-pF.

A sensitivity analysis of the assumptions inherent in the model development has revealed dependencies which are compatible with the above equation and no model *instabilities.*

3.6.2 Performance Limitations

The relative degradation in gate delay is illustrated in Figure 3.34. What is the implication of these predictions ?

They imply that, for a given level of circuit integration and performance, the associated gate delay will be degraded by an amount relative to the delay achievable with an ideal power supply. The amount is related proportionately to the level of circuit integration and circuit performance.

It is clear that any degradation in gate delay may potentially introduce logical errors in any digital circuit and it may therefore be concluded that, if errors are to be avoided, a degradation in circuit performance must be sustained apropos the predicted degradation in gate delay.

Armed with this assertion, the data in Figure 3.34 can be used to assess the performance limit for each circuit size. This was done for the 20MHz and 30MHz degradation data and the average performance limit was calculated. The results are shown graphically in Figure 3.52.

Similar trends are predicted for each of the various non-standard power distribution technologies examined during the course of this study. These are shown graphically in Figure 3.53.

In short, objective (3), outlined in section 3.1, is met by these Figures. Within the technology-based constraints of this analysis, this data represents the extent to which power distribution will limit the integration level and performance of systolic array integrated circuits.

Can the developed simulation methodology be applied to a less regular non-array based digital signal processing architecture with separate constituent processor, memory and control blocks ? Can these constituent circuit blocks be modelled separately and then combined to assess the cumulative effect on the power distribution network ?

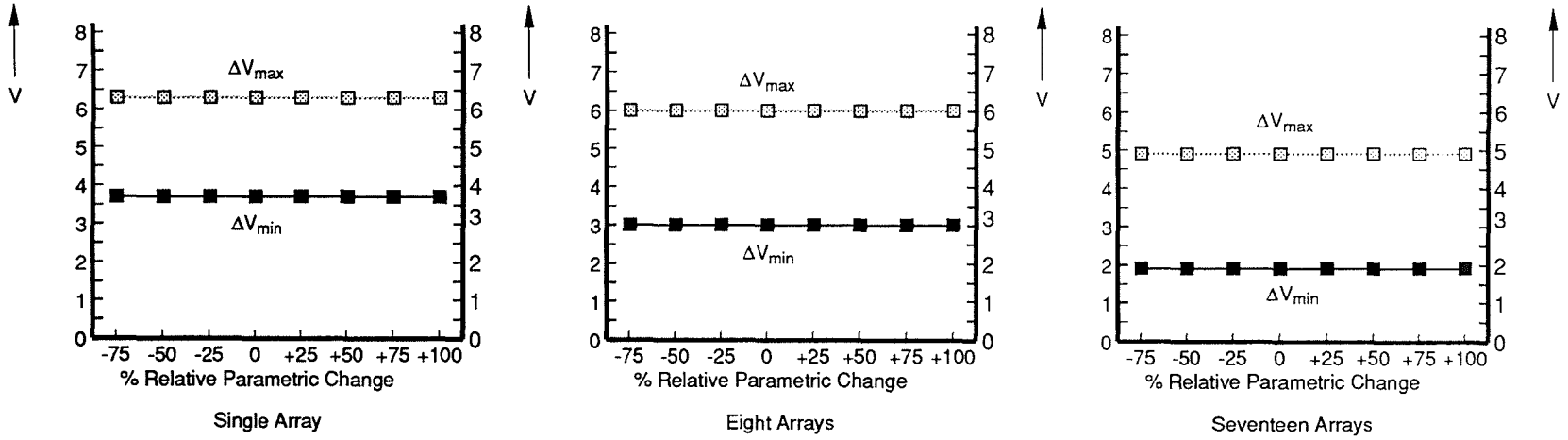


Fig 3.35 Circuit Sensitivity to Bond Wire Resistance

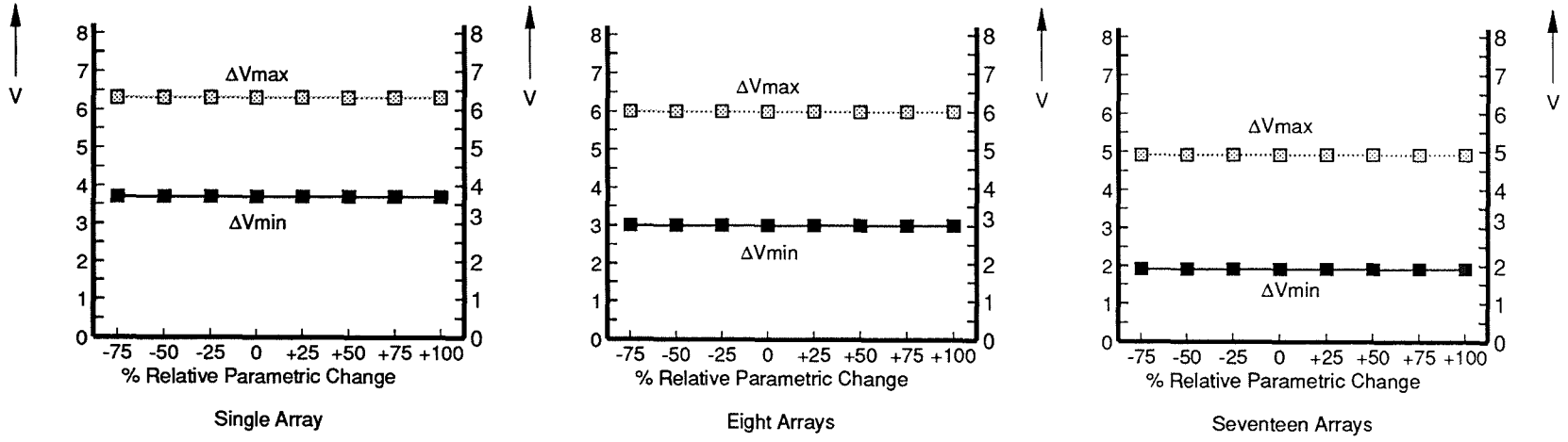


Fig 3.36 Circuit Sensitivity to Bond Wire Capacitance

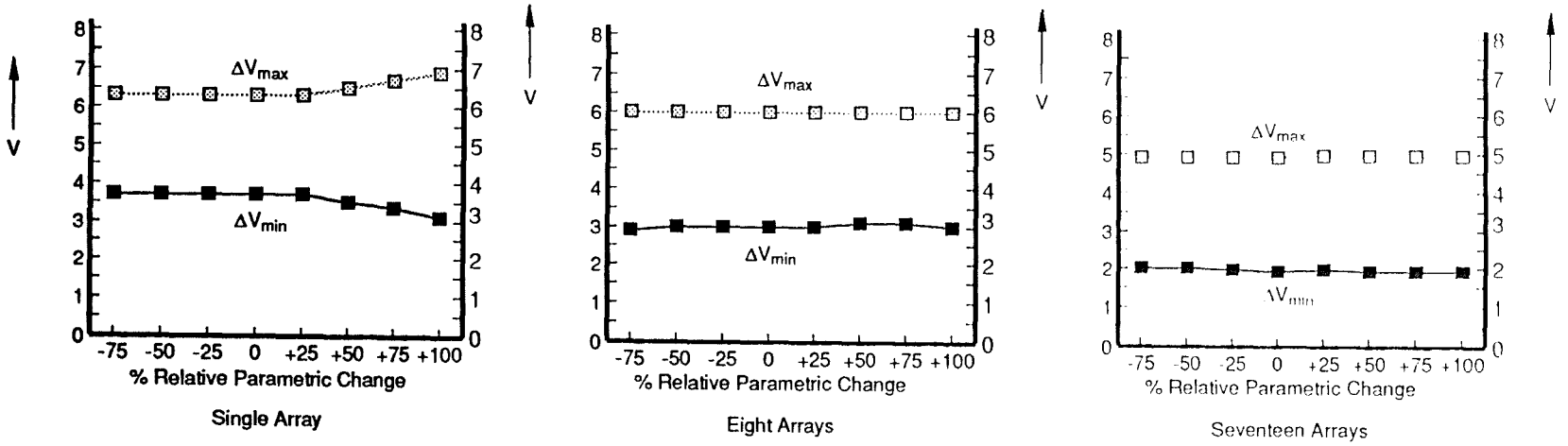


Fig 3.37 Circuit Sensitivity to Bond Wire Inductance

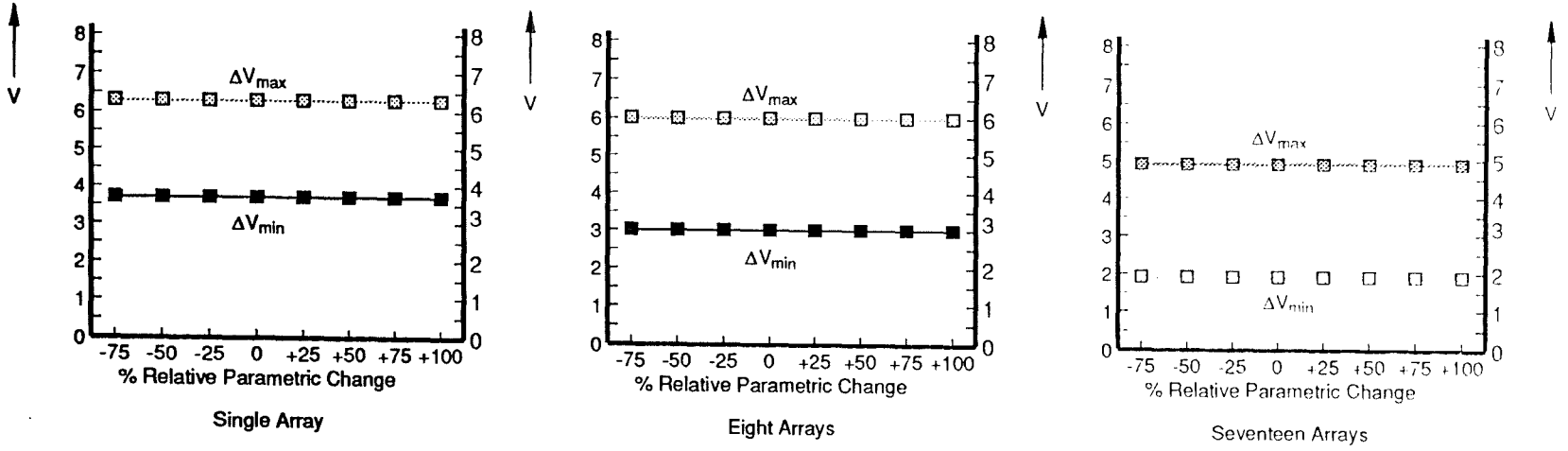


Fig 3.38 Circuit Sensitivity to Pad Capacitance

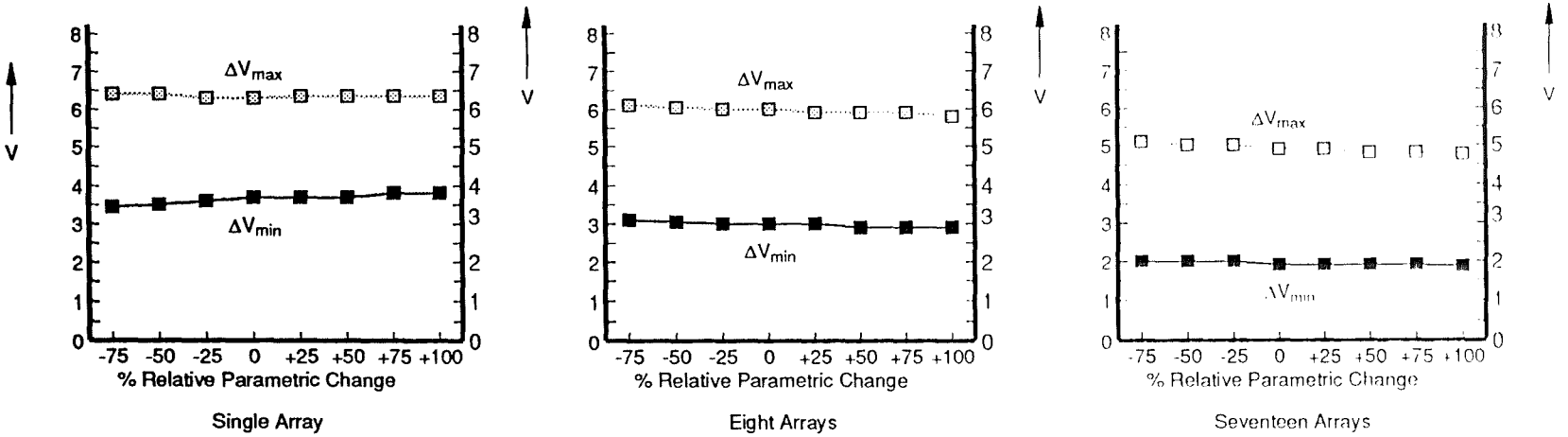


Fig 3.39 Circuit Sensitivity to Package Resistance

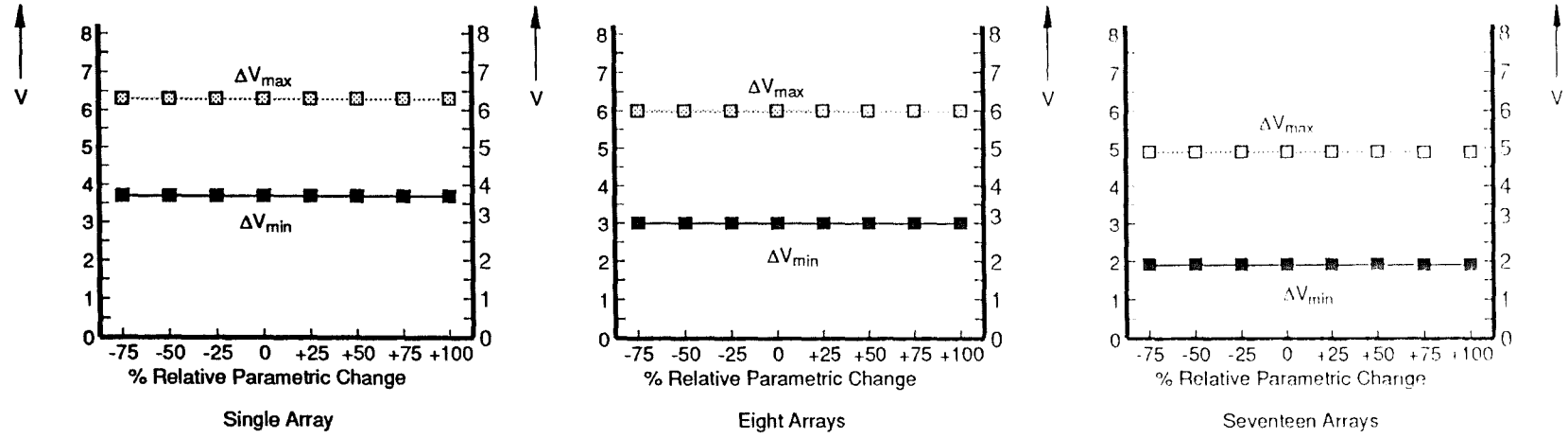


Fig 3.40 Circuit Sensitivity to Package Capacitance

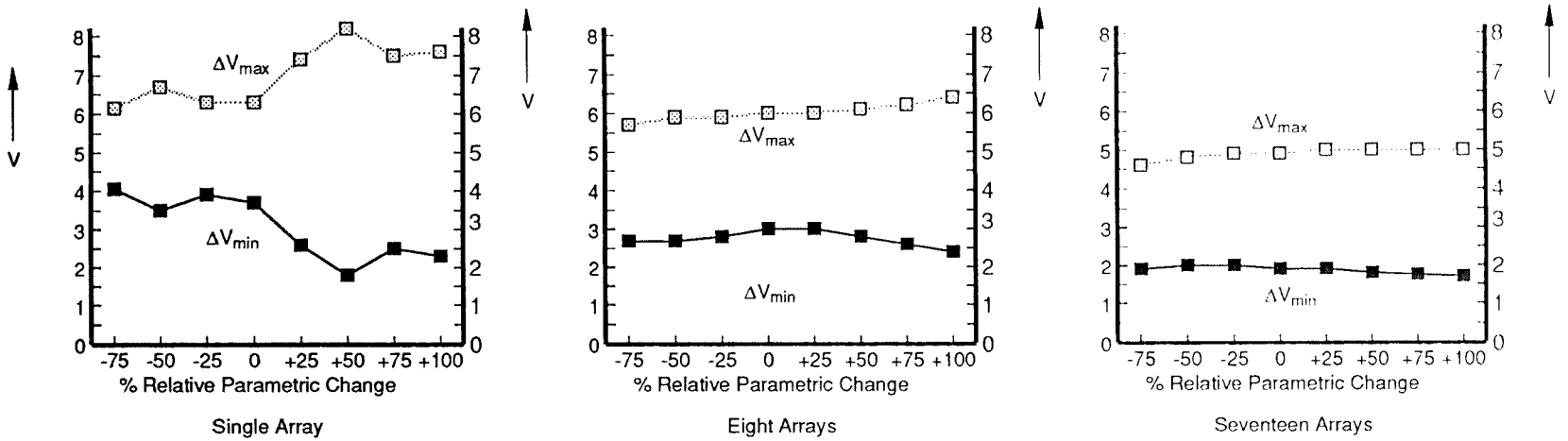


Fig 3.41 Circuit Sensitivity to Package Inductance

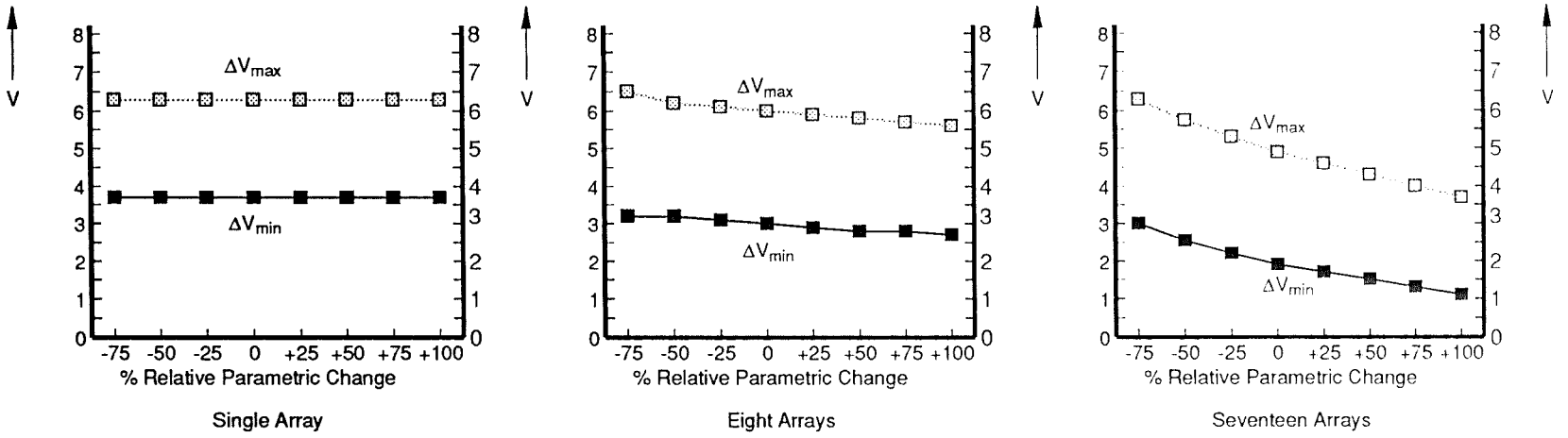


Fig 3.42 Circuit Sensitivity to Network Resistance

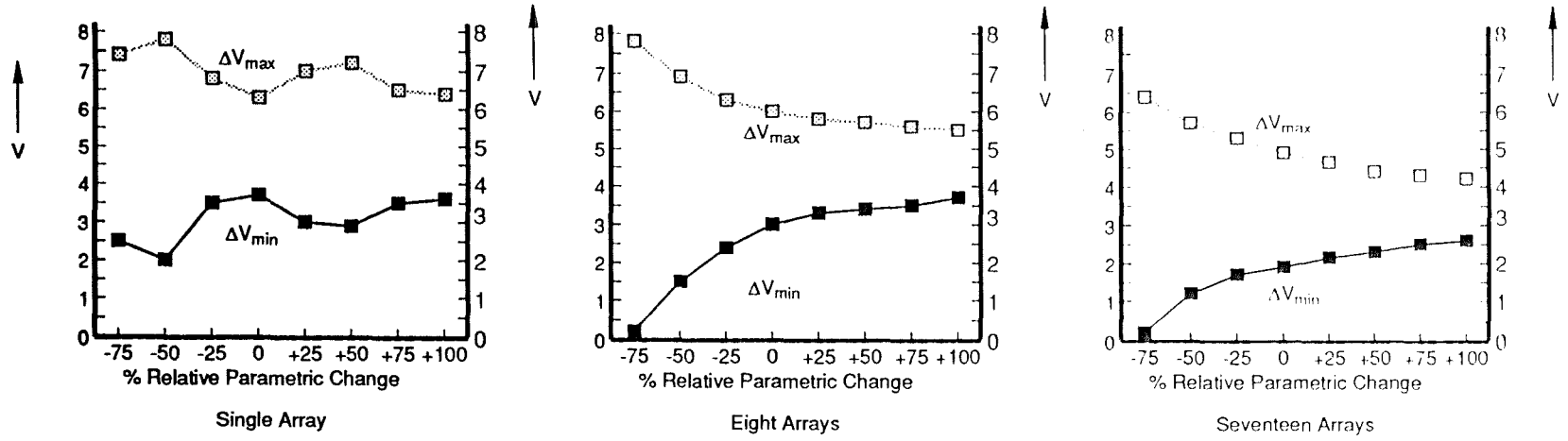


Fig 3.43 Circuit Sensitivity to Network Capacitance

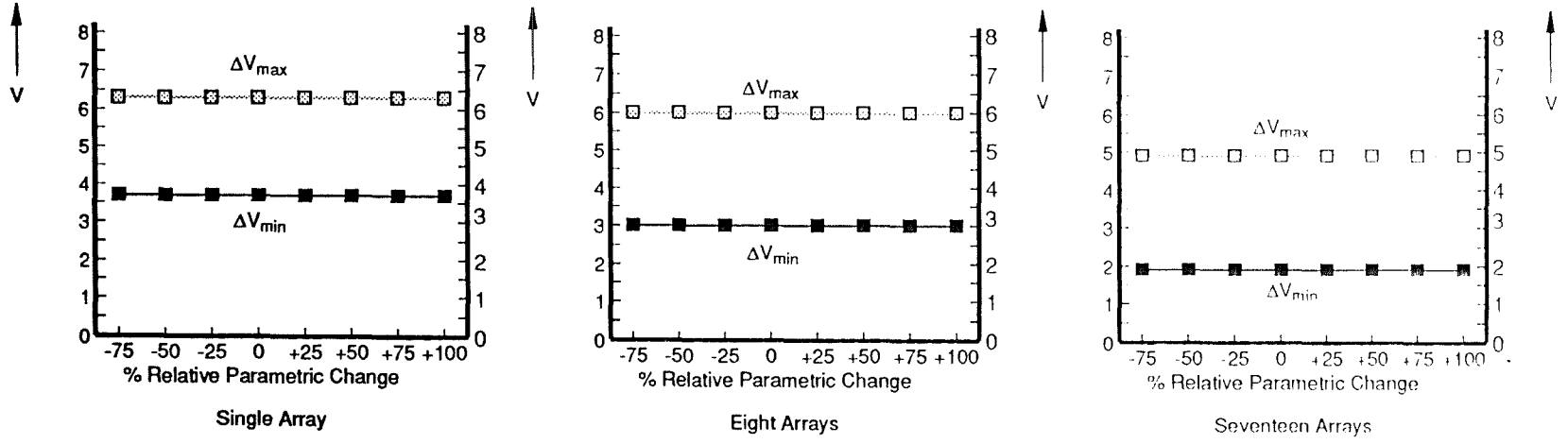


Fig 3.44 Circuit Sensitivity to Network Inductance

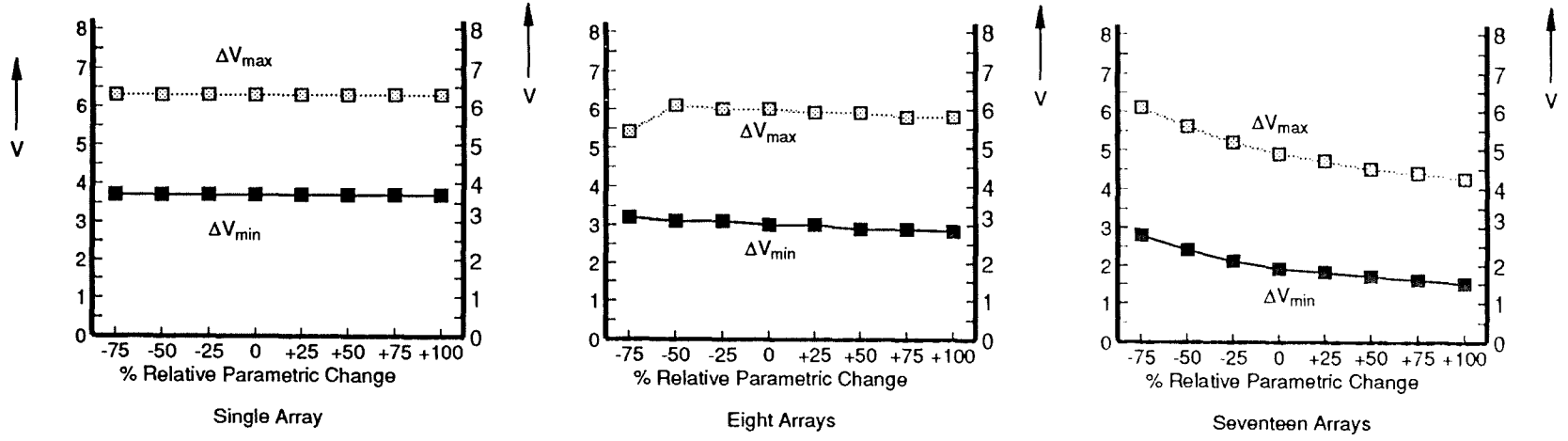


Fig 3.45 Circuit Sensitivity to Metal 1 Resistance

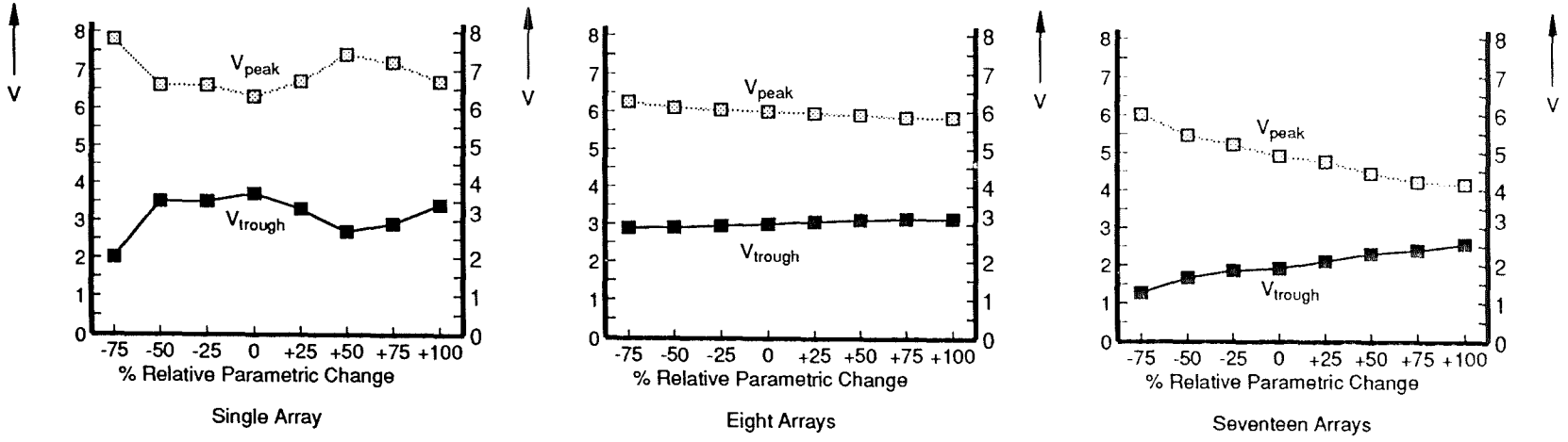


Fig 3.46 Circuit Sensitivity to Metal-1 Capacitance

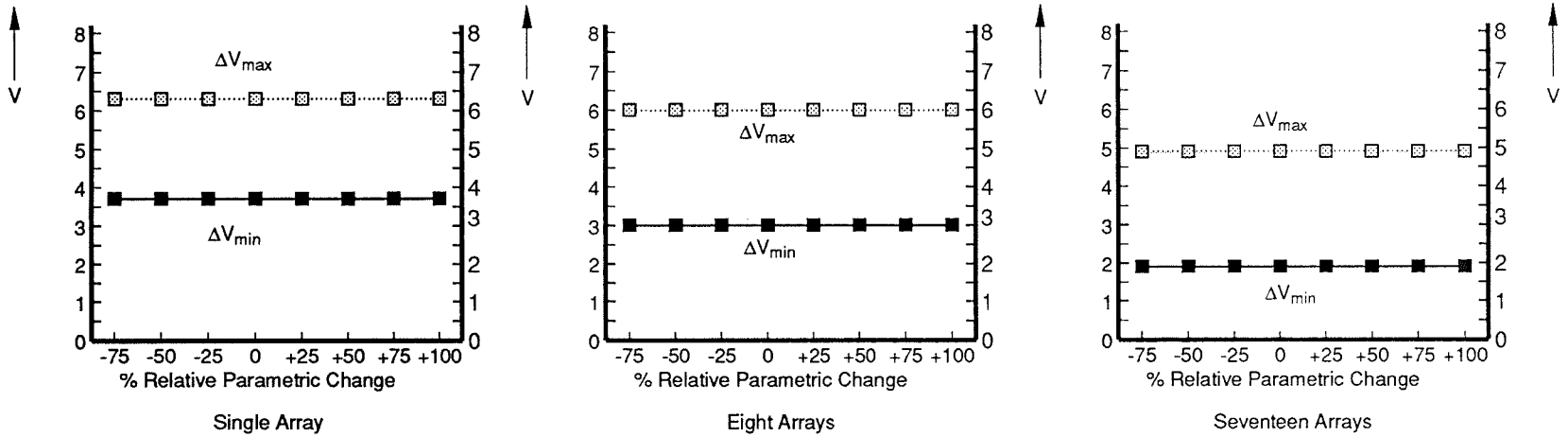


Fig 3.47 Circuit Sensitivity to Metal 1 Inductance

	8 Arrays	17 Arrays
M1	350mV/400mV	1.3mV/1.85V
M1 + M2	500mV/900mV	1.9mV/2.60V
M1/(M1 + M2)	0.7/0.45	0.68/0.71

Fig 3.48 Relative Circuit Sensitivity to Metal 1/2 Resistance

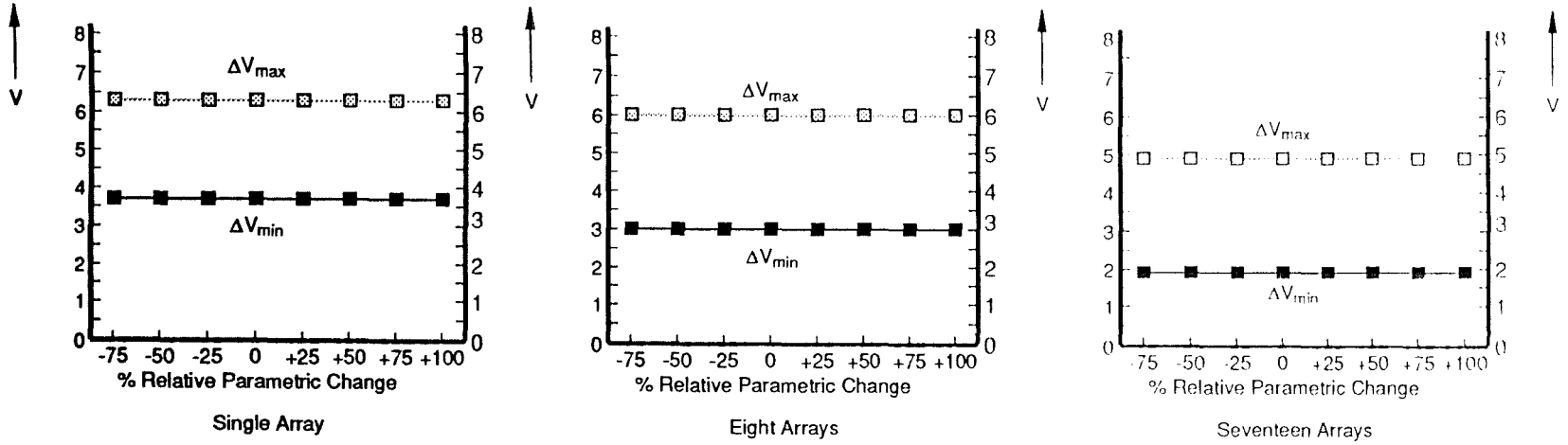


Fig 3.49 Circuit Sensitivity to M1/M2 Contact Resistance

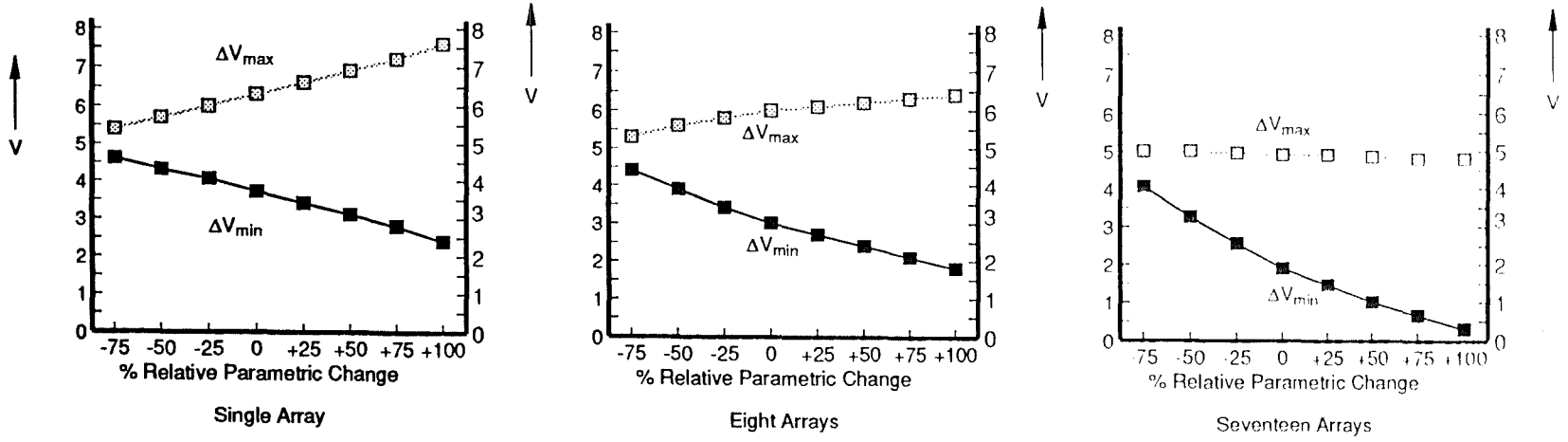
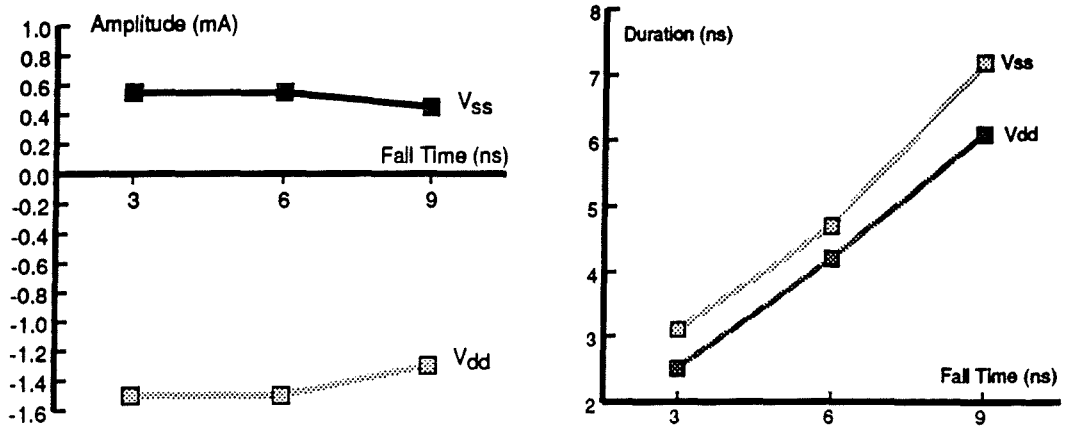


Fig 3.50 Circuit Sensitivity to Load



(a)Pulse Amplitude vs Fall Time **(b) Pulse Duration vs Fall Time**

Fig 3.51 Sensitivity to Test Vector Transition Times

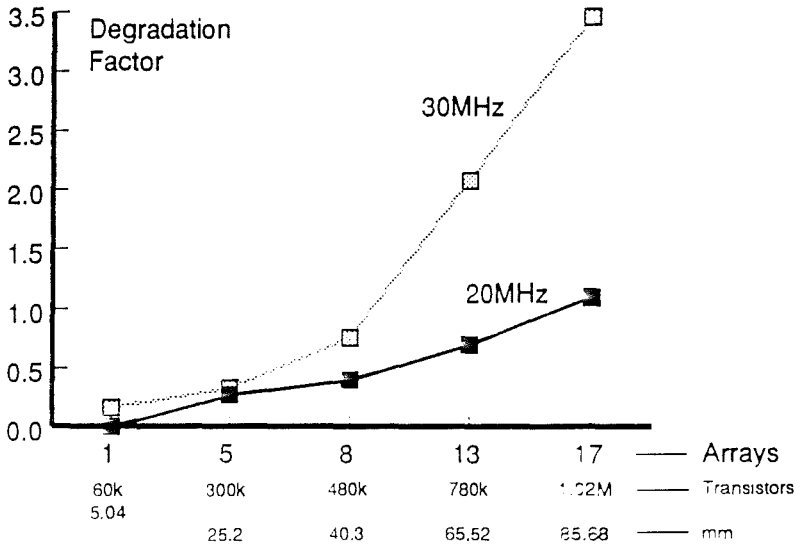


Fig 3.52 Fall Time Degradation vs Circuit Size

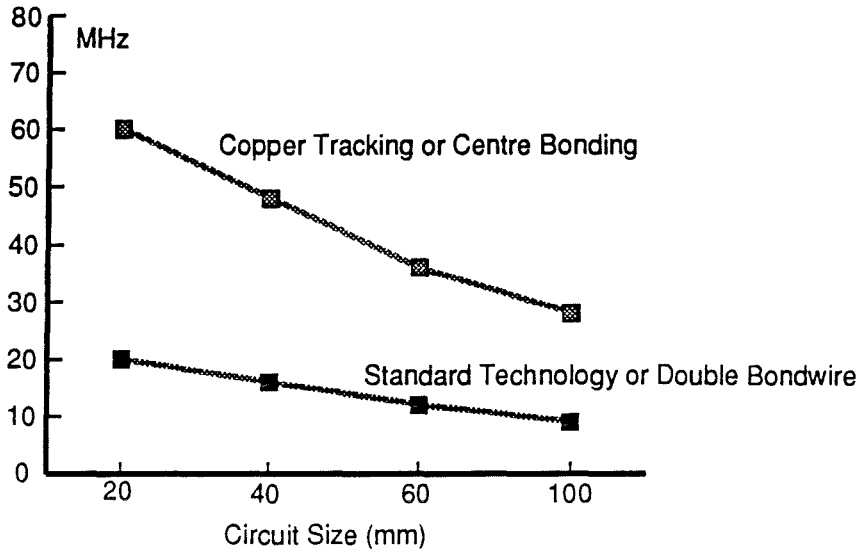


Fig 3.53 Maximum Frequency for Reliable Operation vs Circuit Size

4 POWER DISTRIBUTION NOISE IN A NON-ARRAY-BASED ARCHITECTURE

4.1 Introduction

The objectives of this chapter are (1), to introduce a non-array-based architecture as a vehicle for the analysis; (2), to extend the simulation techniques of chapter 3, thereby to predict power distribution noise for the packaged circuit; and (3), to derive the extent to which power distribution noise limits circuit performance.

In chapter three power distribution noise associated with the systolic array core was assessed. The decision to neglect any noise contribution associated with peripheral control circuits was justified on the grounds that their contribution was insignificant. In the case of a non-array-based architecture, however, it is unjustifiable to assume that any single circuit makes a dominant contribution to power distribution noise.

In this case, it is necessary to assess the effects of individual circuits and then to combine these in order that their net effect be established. Access to circuit-level design information, for the chosen architecture, has made possible such an analysis.

The following structure has been adopted for this chapter:

Section 4.2 is a functional overview of the architecture.

Section 4.3 is a description of each of the three sub-sections which constitute the architecture processor; these are the multiplier unit and the two arithmetic units. This description will be at the circuit level and is needed to understand the development of the processor distribution noise model.

Section 4.4 is a description of each of the various circuits which constitute the device memory blocks. This description will be at the functional level and is needed to understand the development of the memory power distribution noise model.

Section 4.5 is a description of each of the remaining circuit blocks which constitute the device; these are: (1), the ROM blocks; (2), the control blocks; and (3), the I/O blocks. This description will be at the functional level and is needed to understand the power distribution noise models.

Section 4.6 will explain: (1), how each of the noise sub-models for the constituent circuits associated with the processor, were combined to model the whole processor; and (2), how the results emerging from this model were used to assess performance implications.

Section 4.7 will repeat the analysis of section 4.6 for the combined operation of the RAM and ROM memory, control and I/O blocks.

Section 4.8 will draw together and summarise each of the conclusions for sections 4.3 to 4.7. This section is in two parts: part 1 will concern noise modelling; and, part 2 will concern the implications for performance.

4.2 Architectural Overview

The analysis will concern a fault-tolerant signal processor developed by Stewart [e01] to operate with 32-bit floating point data conforming to IEEE standard 754. The design is a dedicated FFT processor and includes sufficient memory to perform a 1024, 512, 256 or 128-point complex FFT. It has been designed to operate at 20MHz using standard 2.0 micron bulk CMOS process technology or at 32MHz using standard 1.5 micron technology. In 2.0 micron technology the circuit will occupy around 6.7 sq.cm and requires around 900,000 transistors. Of these, over 600,000 may be active. The remainder may be used for yield enhancement.

For the purposes of this analysis, the 20MHz version using 2.0 micron bulk CMOS is assumed. The design uses the RADIX-2 algorithm and achieves 200 million floating point operations per second and, in typical arithmetic operations, accesses memory 200 million times per second. This performance is borne out of the unusually high bandwidth available for communication between memory and processor. When in FFT mode, the bandwidth typically is 4×10^{10} bits per second.

A functional block diagram of the device is shown in Figure 4.1 with the associated floor plan in Figure 4.2. Data and inter-rank results are stored in 64k-bits of RAM which is partitioned into 8k-bit blocks to support the communication bandwidth. The FFT coefficients are stored in two identical 8k ROM's. The processing logic consists of two data paths each comprising a 24-bit 40MHz floating point multiplier and two 20MHz floating point arithmetic units. Rounding occurs only as data is written back to the memory thereby allowing maximum accuracy for intermediate computations.

It is clear, from the floor plan of Figure 4.2, that the device may be partitioned, first of all, into processor and memory blocks. Each will be addressed in turn.

4.3 The Processor Block

4.3.1 Block Structure

The overall structure and interconnect topology of the processor block are shown in Figure 4.3. The objectives of the analysis that follows are: (1), to develop a power distribution noise model for the multiplier unit and each arithmetic unit; and (2), to combine these separate models to form a noise model for the "processor slice" shown highlighted in Figure 4.3.

4.3.2 The Multiplier

The multiplier unit is shown in Figure 4.4 from which it is evident that the exponent and mantissa circuitry are mutually independent and consequently may be treated as separate sub-blocks.

4.3.2.1 The Mantissa Sub-block

The 24-bit multiplier, shown in Figure 4.4(a), forms the core of the mantissa sub-

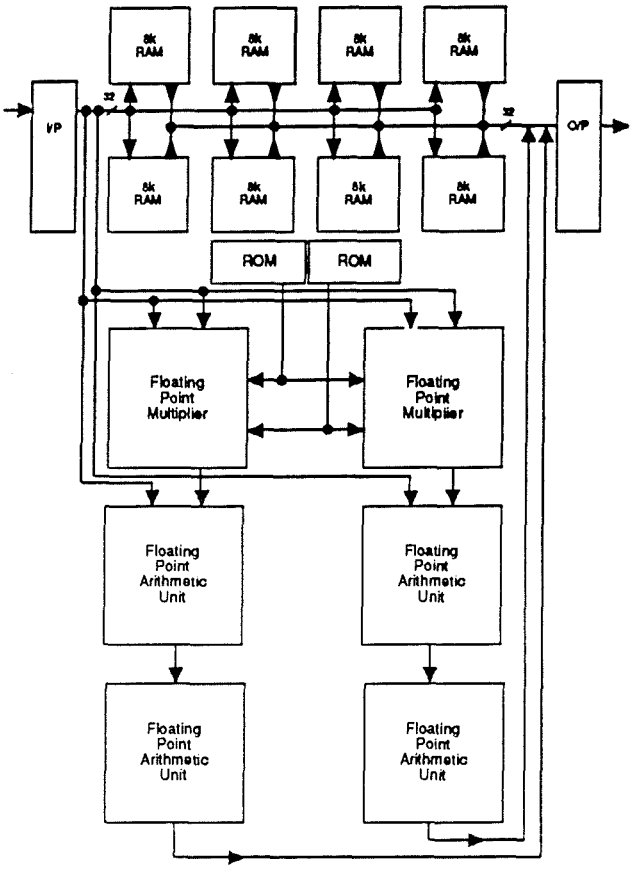


Fig 4.1 Functional Block Diagram

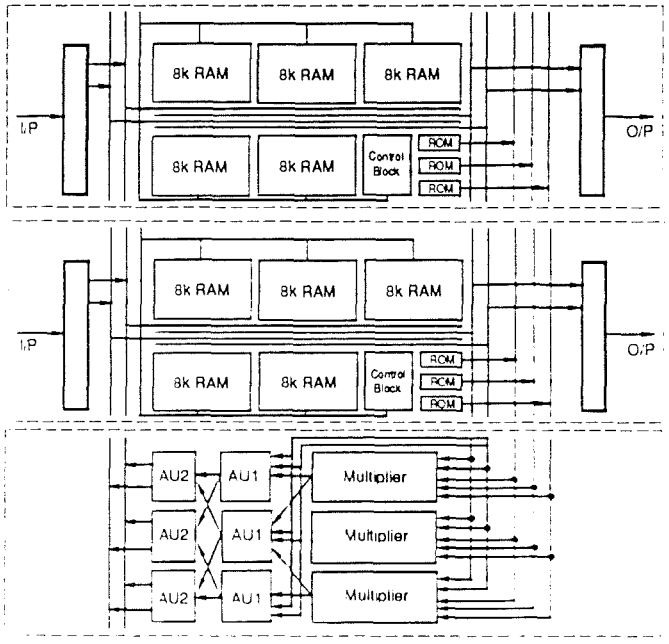


Fig 4.2 Floorplan

block. Each of the 24-bit multiplexers are used to control whether the chosen multiplicands either are real or imaginary.

A representative bitslice, for each of the 24-bit multiplexers, was extracted and simulated at 20MHz; to effect the load associated with the gate to which results are output, a 50fF output capacitor was included. Employing the techniques developed in Chapter 3, the bitslice current flow in each of the positive (Vdd) and negative (Vss) supply lines was captured in separate load-equivalent circuits. In order to model the current flow for the entire 24-bits of each multiplexor, the conductance of the current source in the load-equivalent circuits was scaled appropriately.

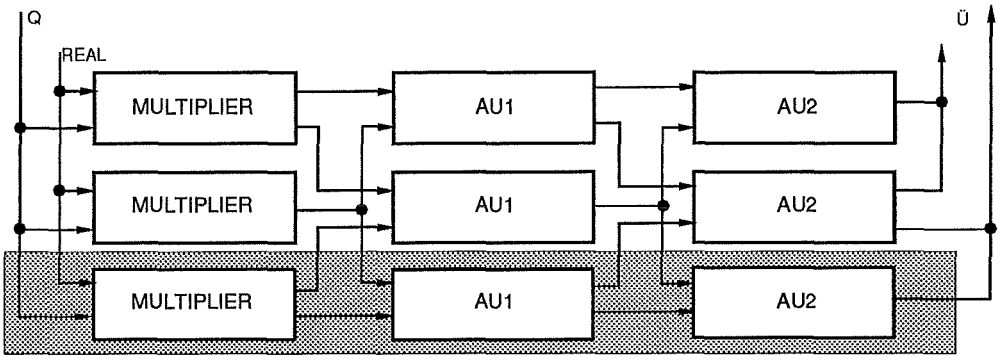


Fig 4.3 Processor - Overall Structure & Interconnect Topology

The internal structure of the 24-bit multiplier is shown in Figure 4.5. It is composed of five "fast register" circuit blocks of which the first four have the structure shown in Figure 4.5; the fifth block is unique.

In the case of the first four fast register blocks, during any particular cycle either real or imaginary data, associated with each multiplicand, are read into the 24-bit latches and stored until, during the subsequent cycle, the data are read into a second latch, 48-bits wide. Each bit associated with one multiplicand, whether real or imaginary, is paired with the appropriate bit from the other, so that, during the third cycle, each of the paired data bits may be input to the carry-save adder. The carry-save adder is 30-bits wide to allow for data-skewing during the multiplication process.

Each of the 24-bit and 48-bit latches was extracted, as a whole, and simulated at 20MHz with 50fF loading. The current flow in each of the positive (Vdd) and negative (Vss) supply lines was captured in separate load-equivalent circuits and, so that the combined effect of both 24-bit multiplicand latches could be modelled, the conductance of the current source in the appropriate load-equivalent circuits was doubled.

In determining the supply line current flow for the carry-save adder, it was recognised that its structure is not regular, thereby making the choice of a representative bitslice element less obvious. After some deliberation, the bitslice element, shown in Figure 4.6, was taken as being representative; it is clear that, in choosing this structure, inherently it is assumed that each bitslice will always generate a carry-bit. This assumption is compatible with the earlier assumption concerning the choice of test

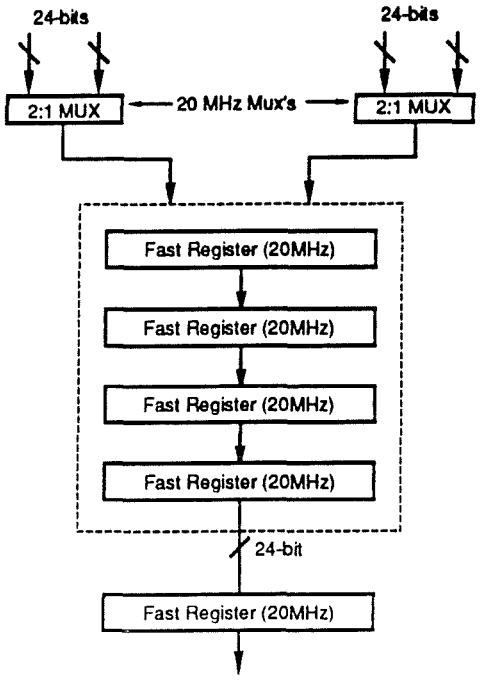


Fig 4.4(a) Multiplier Unit - Exponent Sub-block

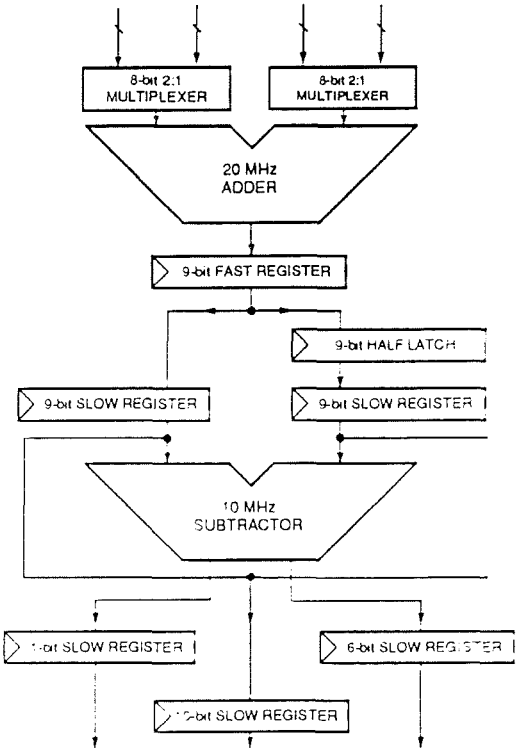


Fig 4.4(b) Multiplier Unit - Mantissa Sub-block

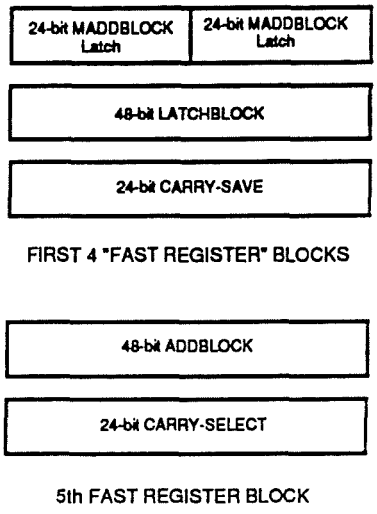


Fig 4.5 Internal Structure of 24-Bit Multiplier

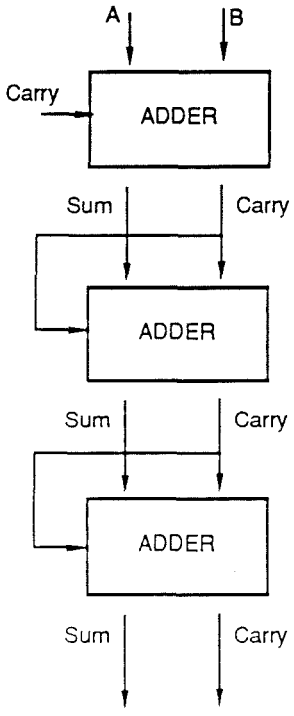


Fig 4.6 Carry-Save Adder "Bitslice"

vectors that are applied during each simulation. It is assumed always that all inputs will effect greatest activity in any given circuit block.

The carry-save bitslice was extracted and simulated at 20MHz with 50fF loading so that the current flow in each of the positive (Vdd) and negative (Vss) supply lines could be captured in separate load-equivalent circuits. The conductance of the current source in each of the load-equivalent circuits was increased thirty-fold in order to model the whole 30-bit adder.

In the fifth and final fast register block, data associated with each multiplicand are read into the 48-bit latch so that each bit, associated with one multiplicand, is paired with the appropriate bit from the other. During the subsequent cycle, these paired bits form the inputs to the carry-select adder which produces a 24-bit output.

The 48-bit latch and the 24-bit carry-select adder were extracted and simulated at 20MHz with 50fF loading. The current flow in each of the positive (Vdd) and negative (Vss) supply lines was captured in separate load-equivalent circuits.

4.3.2.2 The Exponent Sub-block

The circuit diagram of Figure 4.4(b) shows the multiplier exponent circuitry. It is clear that the purpose of the 8-bit multiplexers is to control whether real or imaginary exponent data are added. Addition is implemented with a 9-bit ripple adder and the output is stored in the register shown. Each of these circuit blocks was extracted and simulated at 20MHz with 50fF loading. The current flow in each of the positive (Vdd) and negative (Vss) supply lines was captured in separate load-equivalent circuits.

Next the exponent sum, that is output from the 9-bit 20MHz register, is split into two 9-bit 10MHz data paths. During the first half of a given 10MHz cycle, 9-bit data are output from the 20MHz register to the 10MHz latch on the left data path and, simultaneously, to the 10MHz half-latch on the right data path. At the end of the first half of the 10MHz cycle, therefore, the first data bits to be output from the 20MHz register are held both in the register on the left data path and in the half-latch on the right data path. During the second half of the 10MHz cycle, the new data, output from the 20MHz register, are passed on to overwrite the contents of the register on the left and to those of the half-latch on the right. During this period the original contents of the half-latch also are passed on to be stored in the register on the right data path.

During each 10MHz cycle, two data sets emerge sequentially from the 20MHz register. The first data set are held in the 10MHz register on the left data path and the second data set are held in the 10MHz register on the right data path. The difference between the two exponent sums is computed by the ripple subtractor and is then output, along with the sum of multiplicand exponents, to the first arithmetic unit.

Each of the circuit blocks involved was modelled by modifying the load-equivalent cells developed earlier for the 9-bit adder and for the 20MHz register. In modelling the 20MHz half-latch, since this circuit is similar to the 20MHz register, the load-equivalent circuit for the register was used with the conductance of the current source halved. Similarly, in modelling the 10MHz registers, the same 20MHz 9-bit register equivalent circuit was used with the controlling voltage source modified so that, as in the analysis of chapter 3, the corresponding time separating the current peaks was doubled.

The ripple subtractor was modelled by using the load-equivalent circuit developed for the ripple adder with the conductance of the current source modified to include the additional bit and the voltage source modified to effect operation at 10MHz.

There are three output registers shown in Figure 4.4(b); these are for (1), the 6-bit magnitude of the difference in successive exponent sums; (2), the 1-bit sign of the difference in successive exponent sums; and (3), the 10-bit exponent sum. Each of these was modelled by suitably scaling the conductance in the load-equivalent circuit already developed for the 9-bit 10MHz register which appears earlier in the sub-block.

Having developed load-equivalent circuits for each of the constituent circuit blocks associated with the multiplier unit, it must be recognised that, unlike the systolic array of chapter 3, these blocks do not all become active synchronously; the degree of asynchrony associated with each of the mantissa and exponent circuitry was allowed for as follows.

It may be asserted that the registers, inherently, are synchronous since each is activated by the system clock in the same way as the constituent processors of the systolic array. The degree of asynchrony associated with each constituent circuit block of the multiplier unit, therefore, may be referenced to its associated register; it is the delay of the associated register in combination with the delay associated with all other non-register blocks that appear before it in the pipeline.

For the case of the mantissa circuitry, operating at 20MHz, it is necessary to determine directly the delay associated with the multiplexers, each of the two latches and the carry-save adder, common to the first four multiplier sub-sections, and lastly,

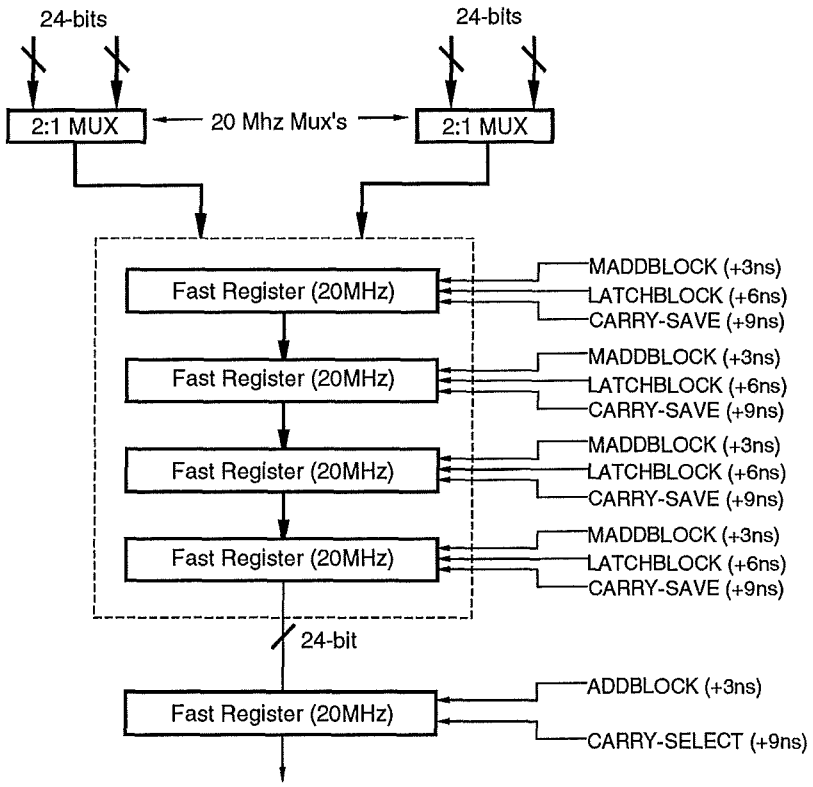


Fig 4.7 Multiplier Unit - Mantissa Sub-block

the latch associated with the final multiplier sub-section. Note that since the carry-select adder of the final sub-section, is placed directly before a register, the delay

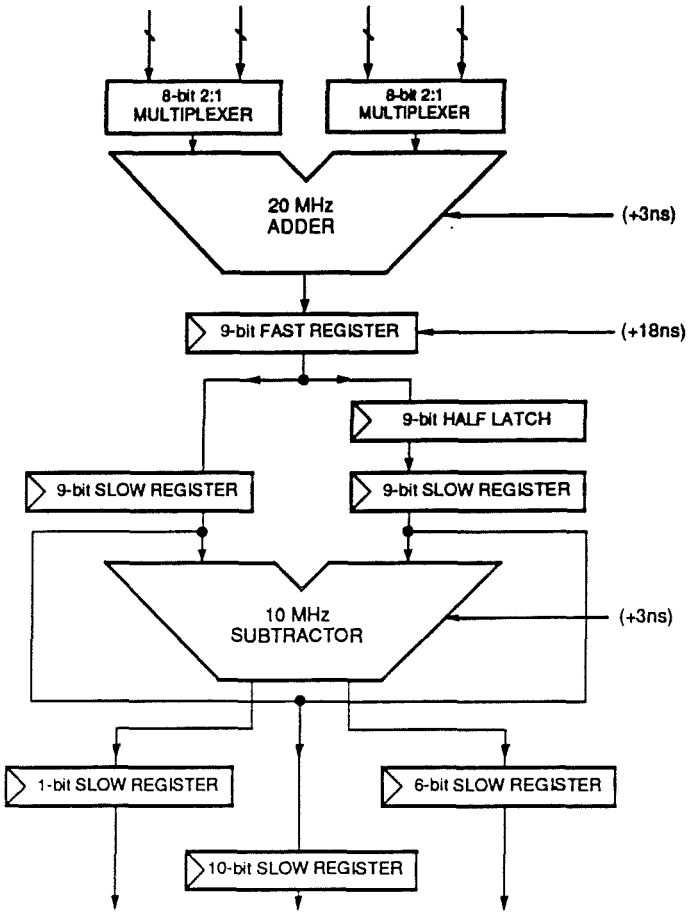


Fig 4.8 Multiplier Unit - Exponent Sub-block

associated with this circuit is irrelevant. Note also that, since the multiplier has a pipelined structure, the delays associated with each sub-section do not accumulate over all five sub-sections. Each of the delays was determined through simulation; the resultant degree of asynchrony for the mantissa block is shown in Figure 4.7.

A similar exercise was undertaken for the exponent circuitry; the result is shown in Figure 4.8.

In order to model this asynchronous circuit activity, the voltage sources for the load-equivalent circuits associated with each non-register block were offset by the delay associated with that block.

4.3.3 The Power Distribution Network

The topology and dimensions of the power distribution network for both the positive (Vdd) and negative (Vss) supply lines are shown in Figure 4.9. Each of the arithmetic units, to be addressed in sections 4.2.3 and 4.2.4, has the same power distribution

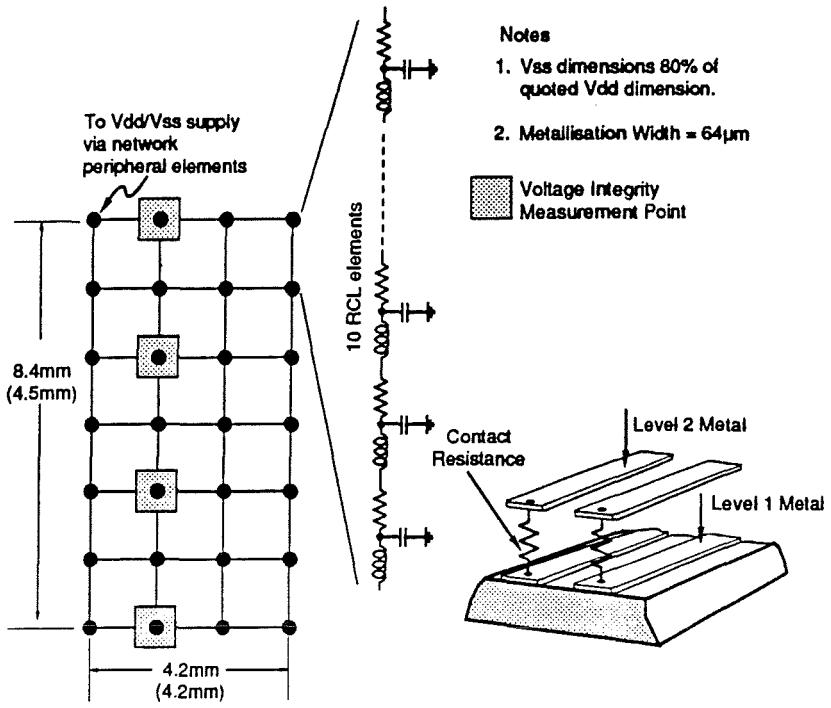
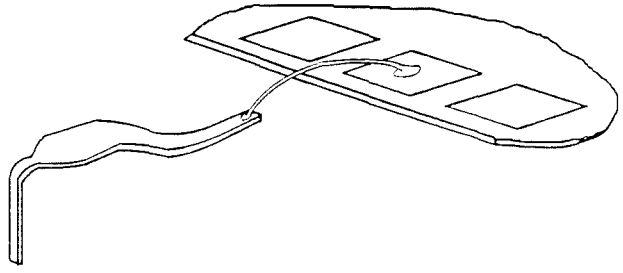


Fig 4.9 Multiplier Power Distribution Topology

topology but with the dimensions shown in parenthesis.

Each section between separate network nodes is modelled by a ten-element RCL ladder network as shown in Figure 4.9. In addition, a two-level metallisation network was developed by substituting equivalent electrical parameters for first and second layer metallisation and choosing inter-network resistance values which correspond to current inter-layer metallisation contact vias used at one hundred micron intervals [d09].



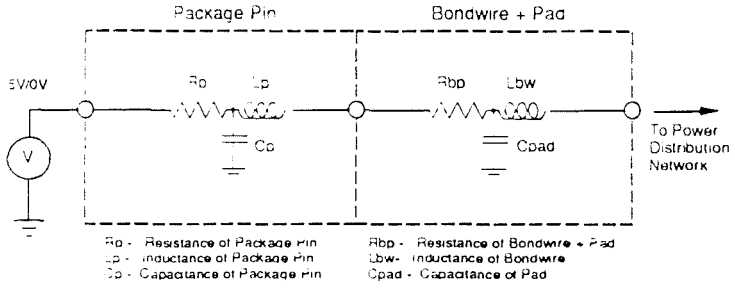


Fig 4.10 Package Related Parasitics

Package-related parasitics were modelled in this case, as they were for the systolic array analysis of chapter 3. Figure 4.10 serves as a reminder of the assumptions made here.

With the load-equivalent model complete and the power distribution network defined, each of the load-equivalent circuits was connected to the first layer metallisation network so that its position corresponded, as closely as possible, to the circuit diagram of Figure 4.4.

4.3.4 The First Arithmetic Unit

The first arithmetic unit is shown in the Figure 4.11 from which it is evident that, here too, the exponent and mantissa circuitry are mutually independent and consequently may be treated as separate sub-blocks in the development of a noise model.

4.3.4.1 The Mantissa Sub-block

With reference to Figure 4.2, it is clear that either successive real or imaginary 49-bit mantissa data, emerging from the multiplier at 20MHz, are loaded into each of the 49-bit 10MHz registers shown in Figure 4.11. This is accomplished using a similar technique to the one used in the exponent circuitry of the multiplier.

During a 10MHz cycle, either real or imaginary 49-bit data emerge in succession from the 20MHz multipliers and are stored in the 49-bit half-latch on the left data stream and in the 49-bit register on the right data stream. During a subsequent 10MHz cycle, data held in the half-latch are released to the 49-bit 10MHz register, while the contents of the half-latch and those of the 10MHz register on the right data stream are updated with new data emerging from the multiplier. At the end of this cycle, therefore, the register on the left data stream holds either real or imaginary data which were released from the multiplier during the initial 10MHz cycle, while the register on the right data stream holds either real or imaginary data which were released during the subsequent 10MHz cycle.

The purpose of the 48-bit 2:2 multiplexer is to load, during each 10MHz cycle, either real or imaginary data from each of the 49-bit registers. The sign bit for the difference in the exponents of these real or imaginary numbers was derived in the exponent circuitry of the multiplier. It is used in the 2:2 multiplexer to determine which of the two mantissa numbers has the smaller exponent. The mantissa with the smaller exponent is output from the multiplexer to the normalisation shifter and the larger to the "function control" block. The single bit which is input to the 2:2 multiplexer immediately preceding the normalisation shifter and function control block, is used to ensure that the mantissa associated with the smaller exponent always is output to the left data stream.

The purpose of the normalisation shifter is to rescale the mantissa associated with the smaller exponent so that effectively it has the same exponent value as the mantissa data with which it is paired. This process of "mantissa normalisation" relies on

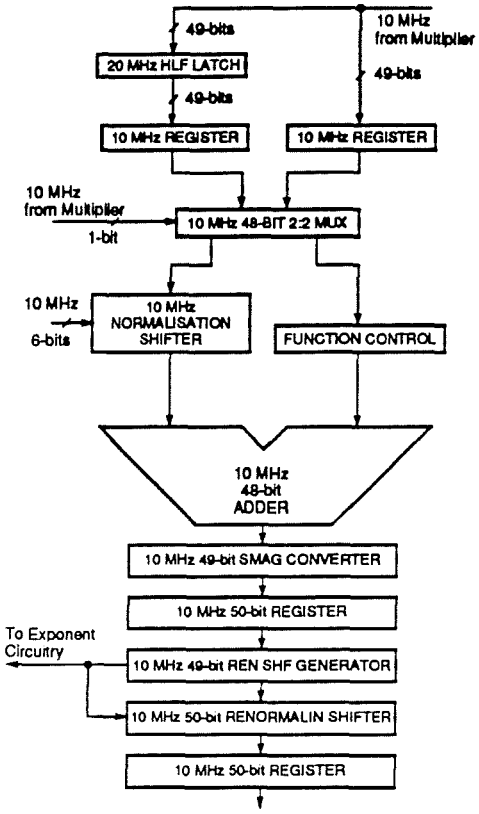


Fig 4.11(a)
First Arithmetic Unit
Exponent Sub-block

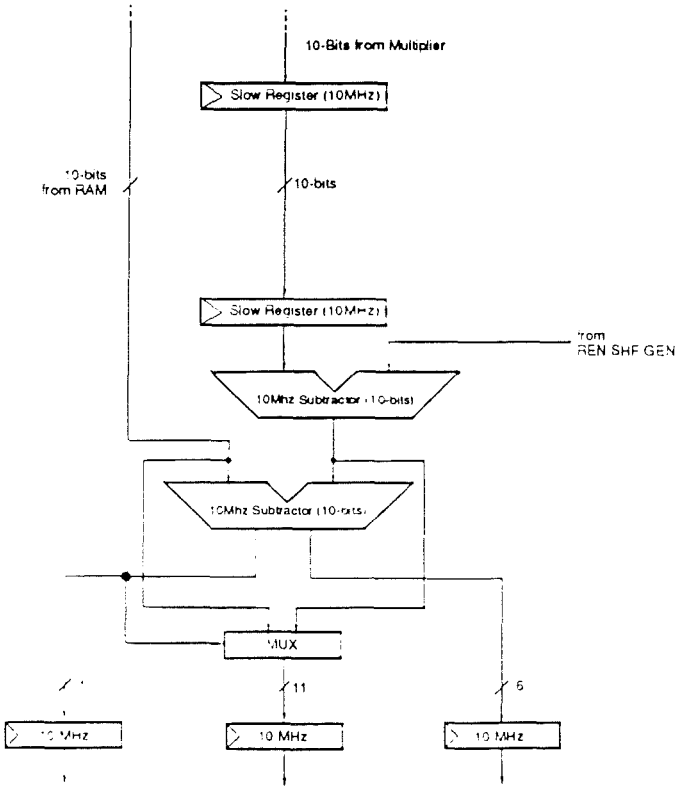


Fig 4.11(b)
First Arithmetic Unit
Mantissa Sub-block

knowing the difference in exponent values as derived in 6-bit form in the exponent circuitry of the multiplier block. The "function control" block simply is used to select whether data associated with the left datastream are subtracted from those of the right, or vice versa, in order that the result which emerges from the adder is positive.

In developing a noise model for the multiplier block, load-equivalent circuits were created for 10MHz 9-bit registers and for a 9-bit 10MHz half-latch. In this case, these load-equivalent models were modified versions of the 10MHz 49-bit registers and for the 49-bit 10MHz half-latch used in the multiplier unit. Each of the 48-bit 2:2 multiplexer, the normalisation shifter and the function control block was extracted and simulated at 10MHz with 50fF loading. The current flow in each of the positive (Vdd) and negative (Vss) supply lines was captured in separate load-equivalent circuits.

The 48-bit carry-select adder, shown in Figure 4.11(a), is used in the addition either of successive real or imaginary mantissa data. The sum which emerges may have to be converted from two's complement format to signed-magnitude format. This is achieved with the smag converter by bit-inversion and adding one to the result which, in signed magnitude form, is stored in the register which follows. The renormalisation shift generator, shown in Figure 4.11, is used to determine the extent to which the signed-magnitude result must be rescaled to produce a normalised number. It does this by counting the number of zeros which appear before a one in the result. The information generated by this block is then used by the normalisation shifter to rescale appropriately the signed magnitude result which subsequently is stored in the register shown.

Since the renormalisation shifter essentially is the same as the normalisation shifter for which a load-equivalent circuit already has been developed, it was necessary to create load-equivalent circuits for the carry-select adder, the signed-magnitude converter and the renormalisation shift generator. In the case of the adder, it was convenient to extract 24-bits and then to double the conductance of the current source associated with the resultant load-equivalent circuit thereby to model the load associated with all 48-bits. The signed-magnitude converter and the renormalisation shift generator were extracted as whole blocks and the current flow in each of the positive (Vdd) and negative (Vss) supply lines was modelled.

4.3.4.2 The Exponent Sub-block

Turning, now, to the exponent circuitry of Figure 4.11(b), one 10-bit input represents the sum of the multiplicand exponents emerging from the multiplier block, while the other 10-bit input is output by the RAM.

Since, in the arithmetic unit mantissa circuitry, the result is renormalised before being output, the sum of multiplicand exponents from the multiplier is required to be reduced by the appropriate degree of renormalisation. This task is performed by subtracting the renormalisation shift from the relevant sum. This is implemented with the 11-bit ripple subtractor shown in the Figure. The 10-bit register has been added to allow adequate time for the renormalisation shift to be generated. The difference

then is found between the sum of exponents associated with the RAM and the corrected sum of exponents associated with the multiplier. This "difference of exponent sums" is used in the normalisation shifter of the second arithmetic unit. Individually, the corrected sum of exponents and the sum associated with the RAM are multiplexed to the 11-bit register where they are output to the exponent circuitry of the second arithmetic unit.

Since load-equivalent circuits, similar to those which are required here, already have been developed for the exponent circuitry of the first arithmetic unit, all that was required was to rescale appropriately the current source conductance for each of the 10MHz registers and the ripple subtractors. The current flow in each of the positive (Vdd) and negative (Vss) supply lines was captured in separate load-equivalent circuits.

Having developed load-equivalent circuits for each of the constituent circuit blocks associated with the first arithmetic unit, the degree of asynchrony associated with each of the mantissa and exponent circuitry was determined as for the multiplier unit.

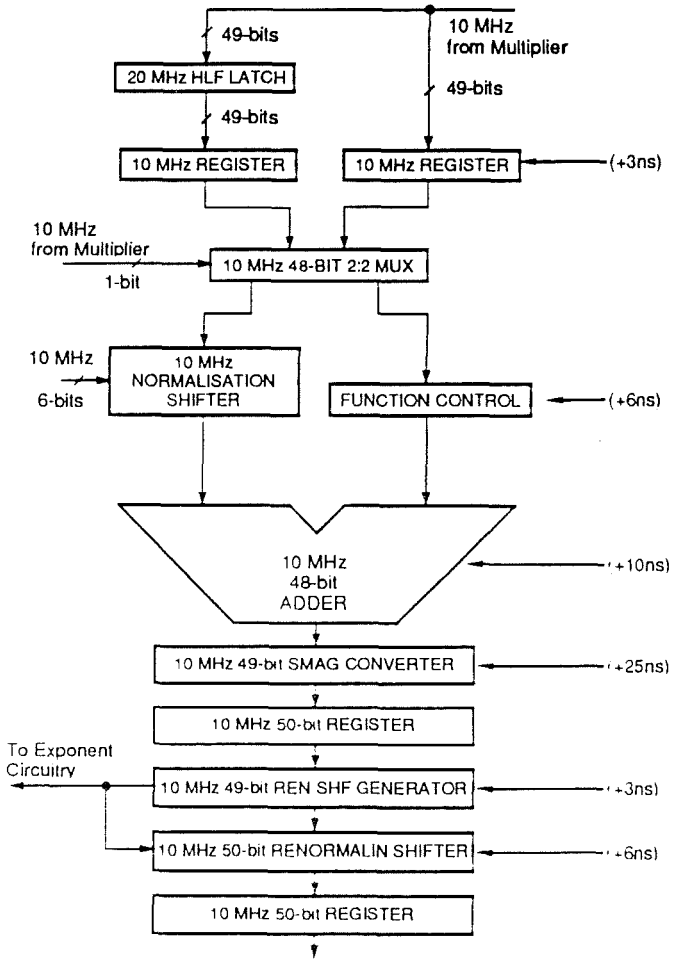


Fig 4.12 First Arithmetic Unit-Mantissa Sub-block

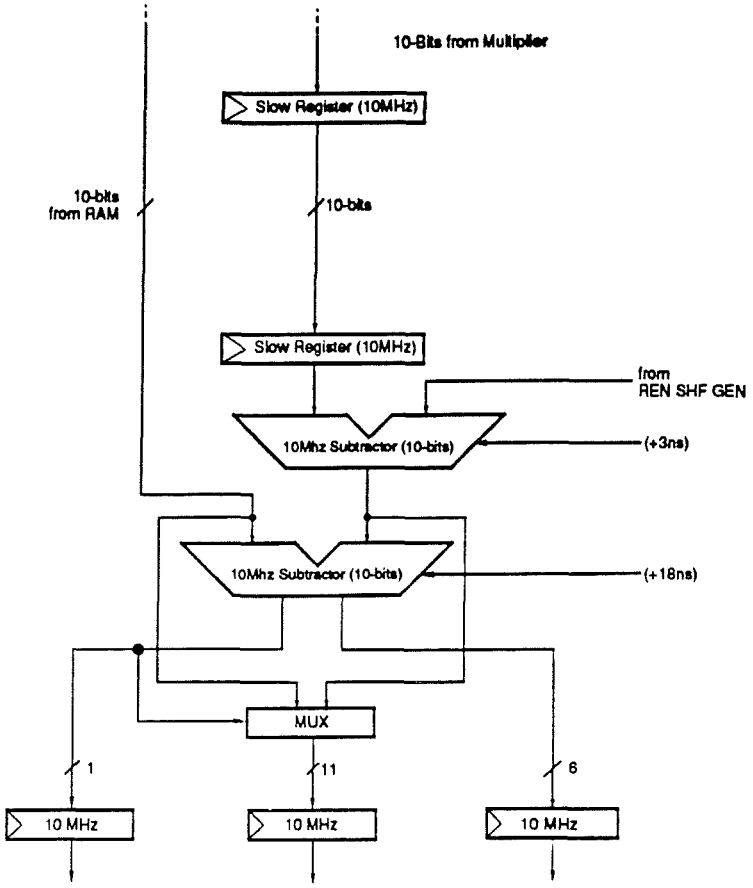


Fig 4.13 First Arithmetic Unit-Exponent Sub-block

For the case of the mantissa circuitry, operating at 10MHz, it was necessary to determine directly the delay associated with the registers, the multiplexor, the normalisation shifter, the carry-select adder and the renormalisation shifter.

Note that the delays associated with the "smag converter" and the renormalisation shifter are irrelevant since each appears directly before a register in the pipeline. Each relevant delay was determined through simulation and the resultant degree of asynchrony is shown in the Figure 4.12. A similar exercise was undertaken for the exponent circuitry; the result is shown in Figure 4.13.

In order to model this asynchronous circuit activity, the voltage sources for the load-equivalent circuits associated with each non-register block were offset by the delay associated with that block.

With the load-equivalent model complete and the power distribution network defined as in section 4.2.2, each of the load-equivalent circuits was connected to the network so that its position in the network corresponded, as closely as possible, to the circuit diagram of Figure 4.11.

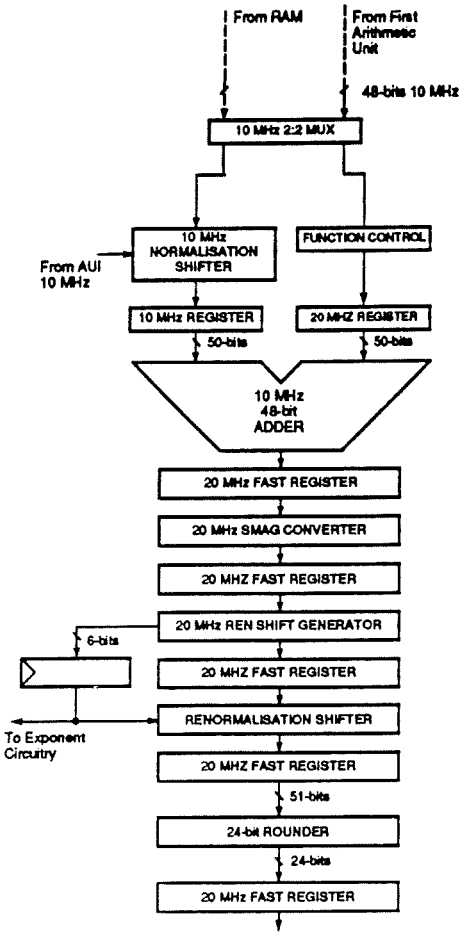


Fig 4.14(a) Second Arithmetic Unit-Mantissa Sub-block

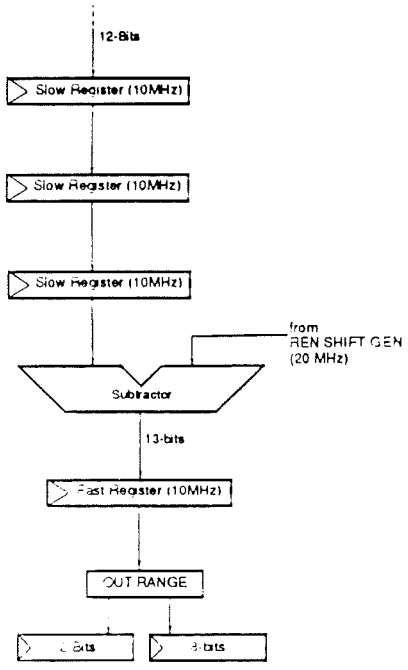


Fig 4.14(b) Second Arithmetic Unit-Exponent Sub-block

4.3.5 The Second Arithmetic Unit

The second arithmetic unit is shown in Figure 4.14. It is evident that the overall structure is similar to that of the first arithmetic unit with most of the constituent circuitry common to both. The main difference with this block is that data are processed at 20MHz and not 10MHz as in the first arithmetic unit.

4.3.5.1 The Mantissa Sub-block

Successive real or imaginary data, emerging from the first arithmetic unit and RAM, are multiplexed by the 10MHz 2:2 multiplexor so that the addend associated with the smaller exponent can be appropriately shifted by the normalisation shifter. As in the first arithmetic unit, the function control block determines whether the left datastream is subtracted from the right, or vice versa, and that data from both are added. The register which follows the function control block therefore must operate at 20MHz while that following the normalisation shifter runs at 10MHz only.

The emerging addends subsequently are added by the 20MHz 49-bit carry-select adder and then are subject to the same processing as in the first arithmetic unit. At the end of the mantissa circuit block, the data are rounded, from 51-bits to 24-bits, before being output by the register shown in Figure 4.14(a).

With the exception of the 24-bit rounder, load-equivalent circuits for each of the circuit blocks associated with the mantissa circuitry of the second arithmetic unit were based on similar circuits already existing for the first arithmetic unit. To effect a different number of bits, the conductance of the current source was scaled appropriately and to effect the change in operating frequency the time separating voltage source peaks was halved.

The 24-bit rounder was extracted and simulated at 20MHz with 50fF loading; the current flow in each of the positive (Vdd) and negative (Vss) supply lines was captured in separate load-equivalent circuits.

4.3.5.2 The Exponent Sub-block

The exponent sub-block for the second arithmetic unit is shown in Figure 4.14(b). It is clear that the "renormalisation shift" is subtracted from the 11-bit data input from the exponent circuitry of the first arithmetic unit and, before the result is output as a 10-bit signed magnitude number, it is compared with the maximum possible exponent to assess whether or not it is within range.

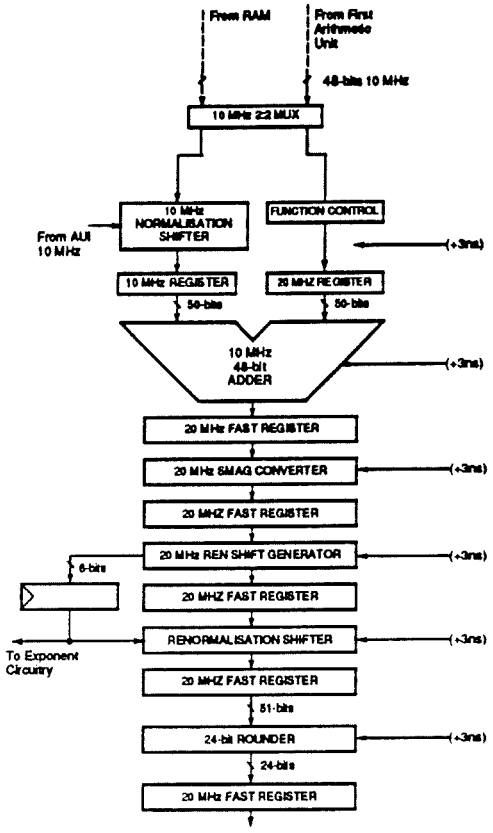


Fig 4.15 Second Arithmetic Unit-Mantissa Sub-block

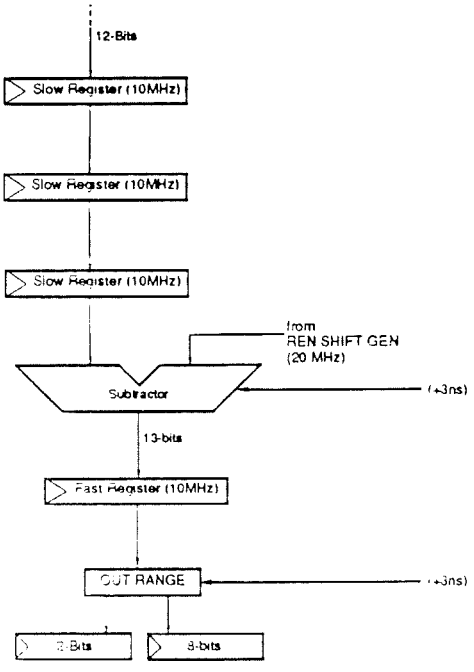


Fig 4.16 Second Arithmetic Unit-Exponent Sub-block

Note that since data from each datastream in the mantissa sub-block are added and subtracted then the renormalisation shifter and subtractor associated with the exponent sub-block must operate at 20MHz.

Note, also, that data are subjected to a growth of one-bit as they progress through the multiplier and arithmetic units. This is associated with the carry-bit which is generated on successive additions and is reconverted to the original width before being output.

Load-equivalent circuits for each of the circuit blocks which constitute the exponent circuitry for the second arithmetic unit were, with the exception of the "out range" block, developed from those of the first arithmetic unit and the multiplier unit by changing appropriately the conductance of the current source and the profile of the controlling voltage source as was done for the mantissa circuit block. The "out range" block was extracted and simulated at 20MHz with 50fF loading. The current flow in each of the positive (Vdd) and negative (Vss) supply lines was captured in separate load-equivalent circuits.

Having developed load-equivalent circuits for each of the constituent circuit blocks associated with the first arithmetic unit, the degree of asynchrony associated with each of the mantissa and exponent circuit was determined as for the multiplier and first arithmetic units.

For the case of the mantissa circuitry, operating at 20MHz, it is necessary to determine directly the delay associated with the registers, the multiplexer, the normalisation shifter and the carry-select adder.

Each relevant delay was determined through simulation and the resultant degree of asynchrony is shown in Figure 4.15. A similar exercise was undertaken for the exponent circuitry; the result is shown in Figure 4.16.

With the load-equivalent model complete and the power distribution network defined as in section 4.2.2, each of the load-equivalent circuits was connected to the network so that its position in the network corresponded, as closely as possible, to the circuit diagram of Figure 4.14.

4.4 The Memory Block

4.4.1 Block Structure

Each of the ten RAM blocks, shown in figure 4.2, is fault-tolerant. Physically, each block is constituted by 10k-bits arranged as four sub-arrays of 68-bits by 39-bits. From these sub-arrays, a total of 8k-bits of functional RAM must be configured. Physically, the RAM has been arranged to be as square as possible so as to minimise wordline and bitline delays. Its physical organisation is shown in Figure 4.17.

To predict the current flow associated with one of the 8k-RAM blocks, each of its constituent circuits was simulated under conditions that are appropriate for 20MHz

operation during both read and write cycles. Simulation conditions appropriate for each constituent circuit will be addressed in the following sub-sections.

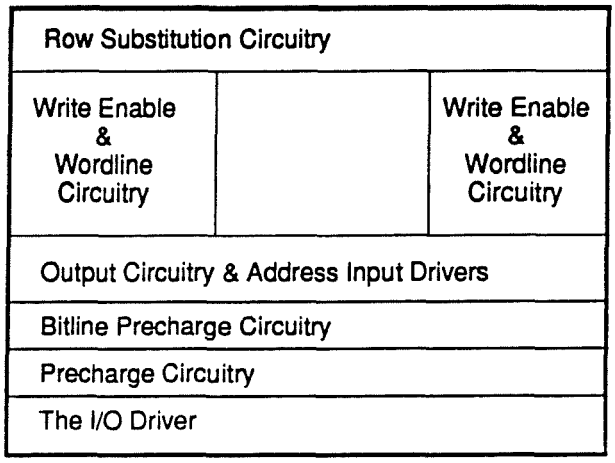


Fig 4.17 8k RAM Block-Physical Organisation

4.4.2 The I/O Driver

The I/O driver is a 32-bit interface for the whole 8k-RAM block. A bitslice element was extracted and simulated at 20MHz with loading that is appropriate for each of the drive directions. For input driving a 50fF load was used and, for output driving, a 1.5pF load was used to represent the interconnect from memory to processor. The current flow in each of the positive (Vdd) and negative (Vss) supply rails was captured in separate load-equivalent circuits. In order to effect the load associated with the complete 32-bit interface, the conductance of the current source in the load-equivalent circuit was scaled appropriately.

4.4.3 The Row Substitution Circuit

Row substitution will occur only during power-up and not during normal circuit operation; the row substitution circuit is not, however, entirely inactive during normal operation.

Each of the four 68-bit by 39-bit sub-arrays has four redundant rows. The row substitution circuit stores the row addresses of up to four rows in as many 7-bit latches. During normal operation, each address which is presented to the substitution circuit is compared with up to four stored values so as to avoid addressing any of the rows which are known to have faults.

The row substitution circuit was extracted and simulated at 10MHz with 50fF loading. The current flow in each of the positive (Vdd) and negative (Vss) supply rails was captured in separate load-equivalent circuits.

4.4.4 The Bitline Precharge Drivers

A bitline precharge driver was extracted and simulated during 10MHz read and write

cycles with 2.5pF loading to represent the combined capacitance of the true and complement bitlines. The current flow in each of the positive (Vdd) and negative (Vss) supply lines was captured in separate load-equivalent circuits. Remembering that the RAM is organised as four sub-arrays each of 68 rows and 39 columns, it is clear that there will be 156 (4X39) bitline drivers in simultaneous operation. The conductance of the current source was scaled appropriately.

It is clear, from Figure 4.18, that the difference in supply current during the read and the write cycles is significant. During a read cycle, the true and complement bitlines are precharged to about 3.5V and subsequently, as the data are read, this level, on either of the bitlines, is required to change only by the order of a hundred millivolts for the sense amplifiers to be triggered. During a write cycle, however, the bitlines again are precharged to around 3.5V and subsequently, as the data are written, this level on either of the bitlines will be driven either to 5V or to 0V. Clearly, the write cycle will place higher demands on the supply rails.

4.4.5 The Precharge Circuits

Since RAM block precharging is dominated by bitline precharging, as described above, circuit blocks responsible for the remaining RAM precharge tasks were extracted as a single circuit combination. This combination was simulated at 10MHz and with appropriate loading. The current flow in each of the positive (Vdd) and negative (Vss) supply lines was captured in separate load-equivalent circuits.

4.4.6 The Address Input Drivers

The address input drivers are 9-bits wide. Two of these bits, bits 0 and 1, are used to determine the bit-line and the remaining seven, bits 2 to 8, are used to determine the word-line. In any particular cycle, no more than seven of the nine input drivers simultaneously may be active.

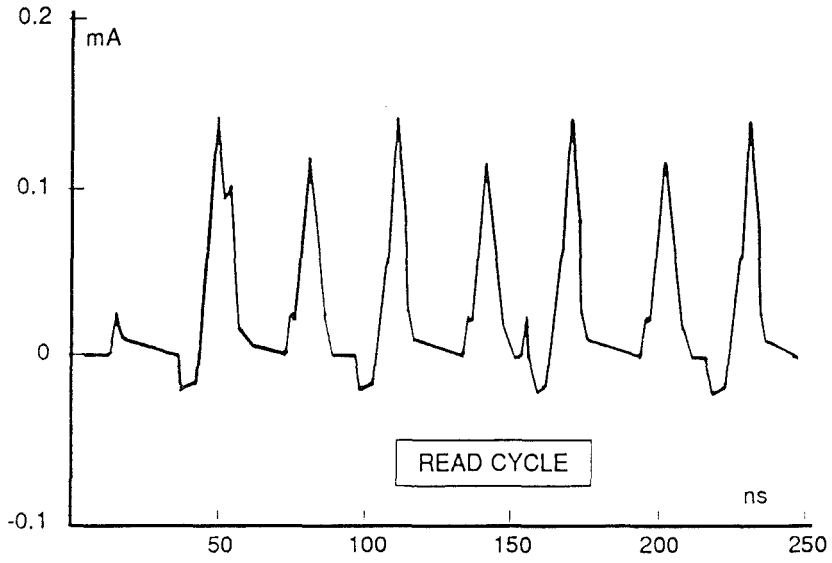


Fig 4.18(a) Vss Bitline Precharge Current- Read Cycle

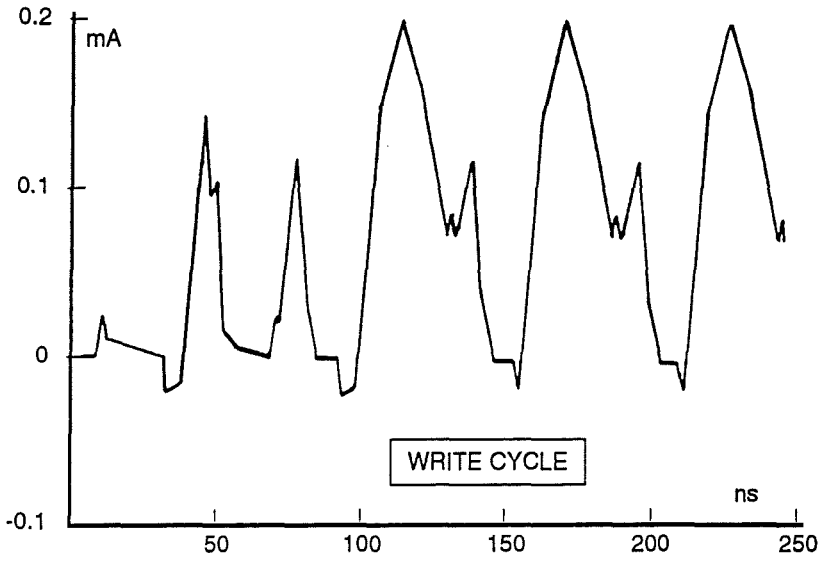


Fig 4.18(b) Vss Bitline Precharge Current - Write Cycle

A single driver was extracted and simulated at 10MHz with appropriate loading. The current flow in each of the positive (Vdd) and negative (Vss) supply lines was captured in separate load-equivalent circuits and, in order to model the effect of seven drivers operating simultaneously, the conductance of the current source was scaled appropriately.

4.4.7 The Output Circuitry

The output circuitry, as distinct from the 32-bit I/O driver, was extracted for each of the 39 columns. Each block was simulated at 10MHz with appropriate loading and a load-equivalent circuit representing the current flow for all 39 columns was created.

4.4.8 The Remaining Circuit Blocks

Having addressed each of the higher dissipating circuit blocks, the remaining blocks, which include the write enable and wordline circuits, were extracted as a single circuit block and simulated at 10MHz with appropriate loading. The current flow in each of the positive (Vdd) and negative (Vss) supply lines for the combined block was captured in separate load-equivalent circuits.

4.4.9 The Power Distribution Network

An RCL ladder network model of the power distribution scheme for the 8k-RAM block was developed. The topology and dimensions of this network are shown in Figure 4.19. As in previous distribution network models each element is made up of ten RCL sub-elements and the distribution network topology is implemented on both metallisation layers interconnected by contact vias.

As shown in Figure 4.18, the current flow for the bitline precharge drivers, during a

write cycle, has been predicted to be around 30% higher than during a corresponding read cycle. In addition, the current flow associated with the 32-bit I/O driver similarly will be higher when it is driving out, as during a read cycle, than when it is driving in. These facts have meant that separate noise models for the RAM block during read and write cycles need to be developed.

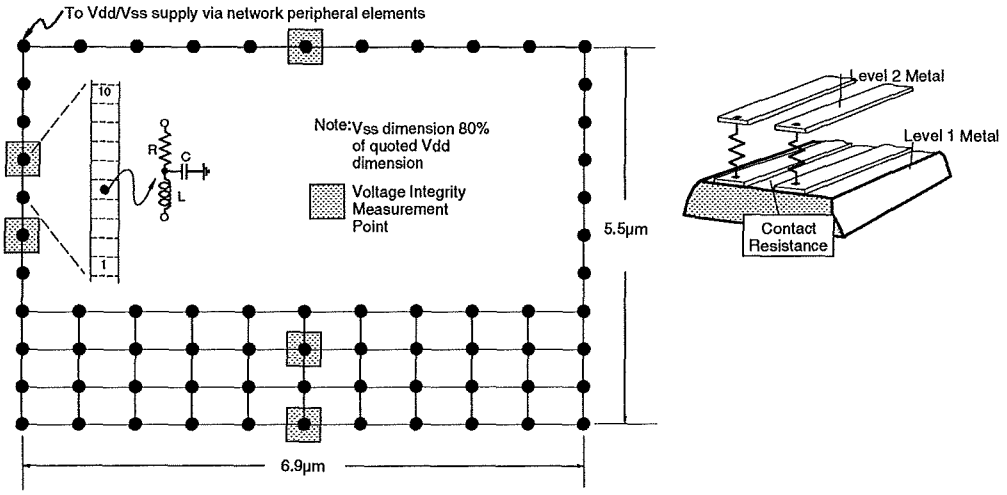


Fig 4.19 Power Distribution Network Topology for Memory Reticle

Each of the load equivalent circuits representing constituent circuit blocks was connected to the distribution network as shown in the Figure 4.19, so as to correspond as closely as possible with the actual 8k-RAM floorplan. As in previous cases, the conductance of the current source associated with each load-equivalent was reduced by a factor of ten so that each equivalent circuit could be evenly distributed over ten distribution network elements.

4.5 The Control, ROM and I/O Blocks

4.5.1 The Control Block

The physical organisation of the control block is shown in Figure 4.20. The purpose of this block is to manage the data flow path between memory and processor blocks. As part of this task, the control block must access only the functional parts of each of the fault-tolerant RAM blocks. Within the control block this type of address generation is achieved with the address sequencer and address translator blocks. These blocks constitute much the greatest contribution to the control block power supply noise since, during normal operation, the "signature" and "configure" blocks remain inactive.

During normal operation, the address sequencer generates the address locations which actually are required, while the address translator translates these addresses to those corresponding addresses which have been found by mapping to the reconfigured RAM. The end result is that each requested address location is never mapped on to a faulty address location.

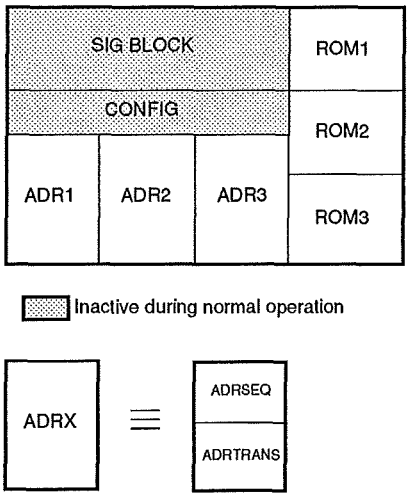


Fig 4.20 Control Block - Physical Organisation

The address sequencer and address translator circuit blocks were extracted and simulated at 10MHz with 8pF loading to represent the rather extensive interconnect from the control block to each of the RAM blocks. The current flow in each of the positive (Vdd) and negative (Vss) supply lines was captured in separate load-equivalent circuits.

4.5.2 The ROM blocks

Two identical 8k-ROM's are used to store the FFT coefficients. A ROM block was extracted and simulated at 10MHz with appropriate loading. The current flow in each of the positive (Vdd) and negative (Vss) supply lines during a single cycle was captured in separate load-equivalent circuits.

4.5.3 The I/O Blocks

A representative bitslice for each of the 32-bit read and write ports, as are shown in Figure 4.2, was extracted and simulated at 20MHz with 5pF loading to represent the interconnect from write port to memory and 12pF loading to represent the load experienced when driving off-chip as during a read cycle. It is evident, from Figure 4.21(b), that the write port, as designed, does not provide adequate gain. This inevitably will mean that the data must be written at lower rates.

The current flow in each of the positive (Vdd) and negative (Vss) supply lines was captured in separate load-equivalent circuits. To model the effect of 32-bits, the conductance of the current source was scaled appropriately in each case.

4.5.4 The Power Distribution Network

The power distribution network model for the control and ROM block is similar to that of the 8k-RAM block. One control block and two ROM blocks were positioned on the network as shown in Fig 4.22. Here again, the conductance of the current source in each of the load equivalent circuits was reduced so that the more realistic effect on

the network of an array of reduced load equivalent circuits was modelled.

The distribution network for the 32-bit read and write ports is shown in Fig 4.23. This distribution scheme was modelled as in previous cases.

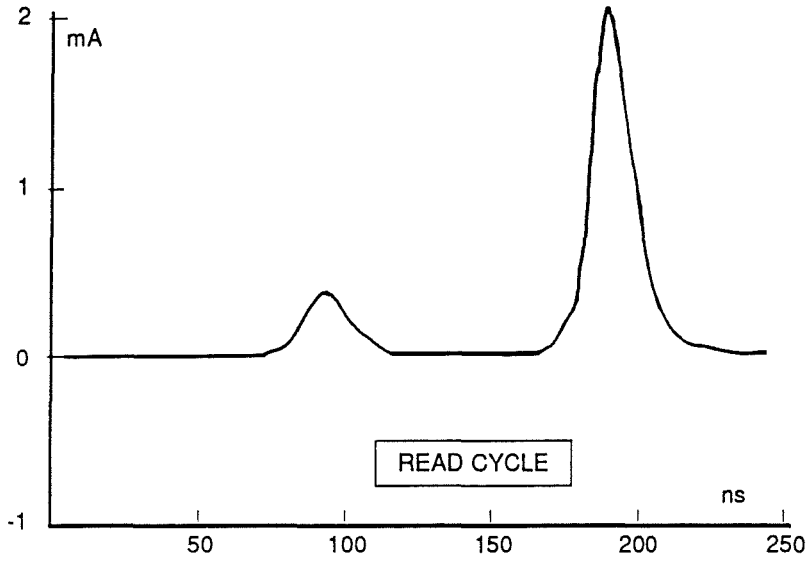


Fig 4.21(a) I/O Port Vss Current-Read Cycle

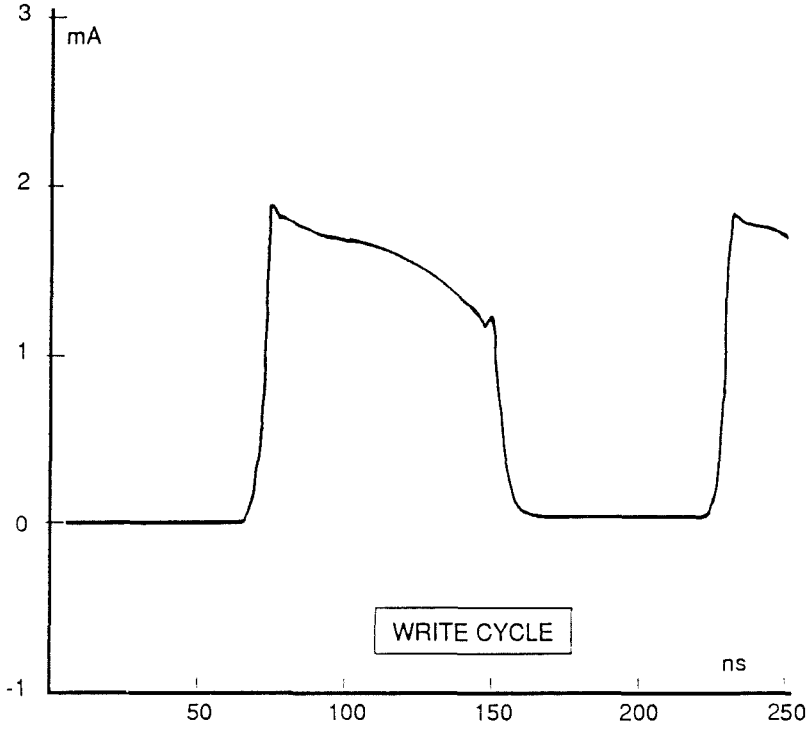


Fig 4.21(b) I/O Port Vss Current-Write Cycle

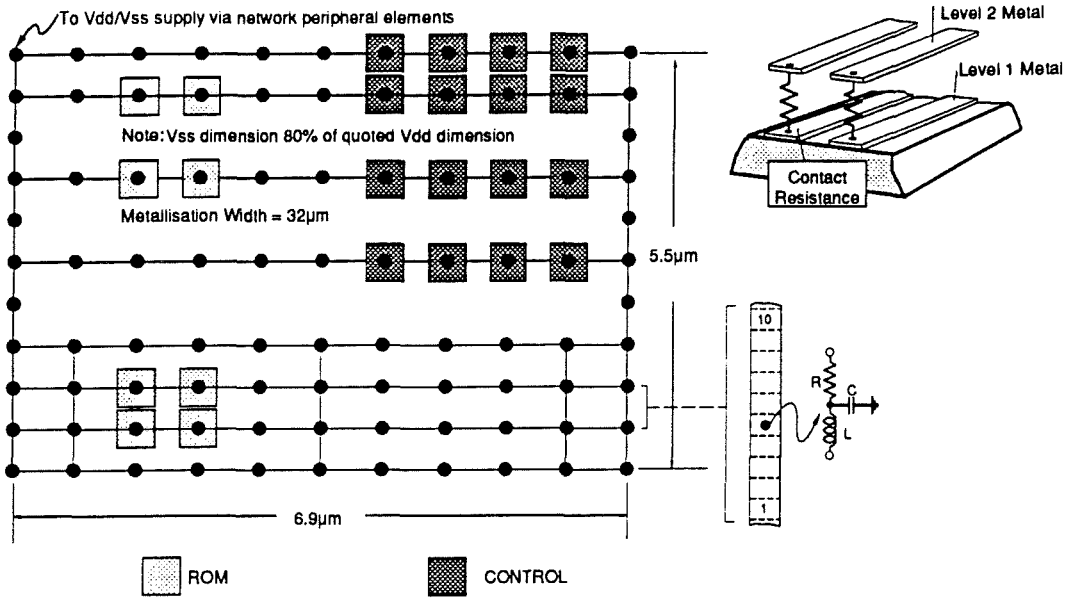


Fig 4.22 Distribution Network for Control Block

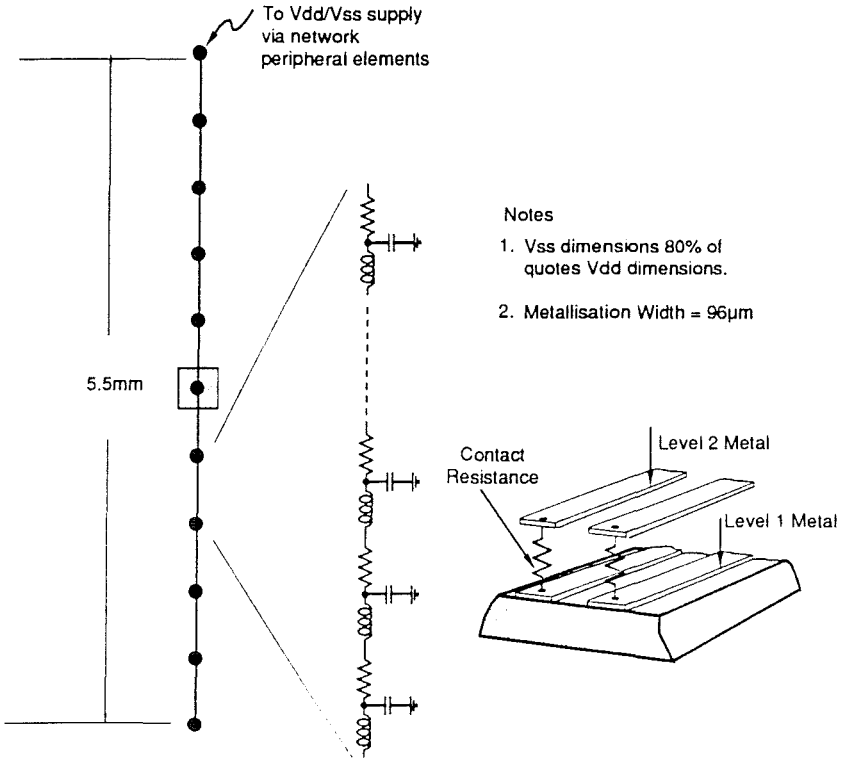


Fig 4.23 Distribution Network for Read & Write Ports

4.6 Performance of Processor Reticle Power Distribution Network

4.6.1 Noise Model

Noise models for the multiplier unit and each of the arithmetic units were combined to form a noise model for the entire "processor slice" shown highlighted in Figure 4.3. Power supply connections, shown in Figure 4.24, were arranged so as to correspond with the prototype device.

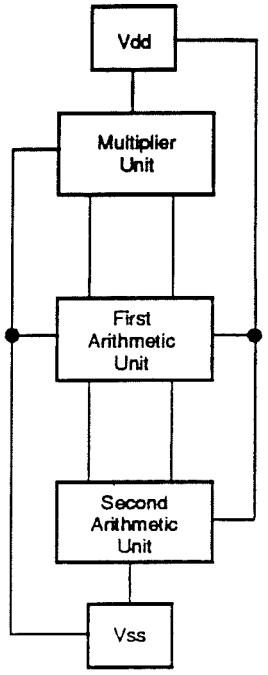


Fig 4.24 Power Distribution Network for Processor Slice

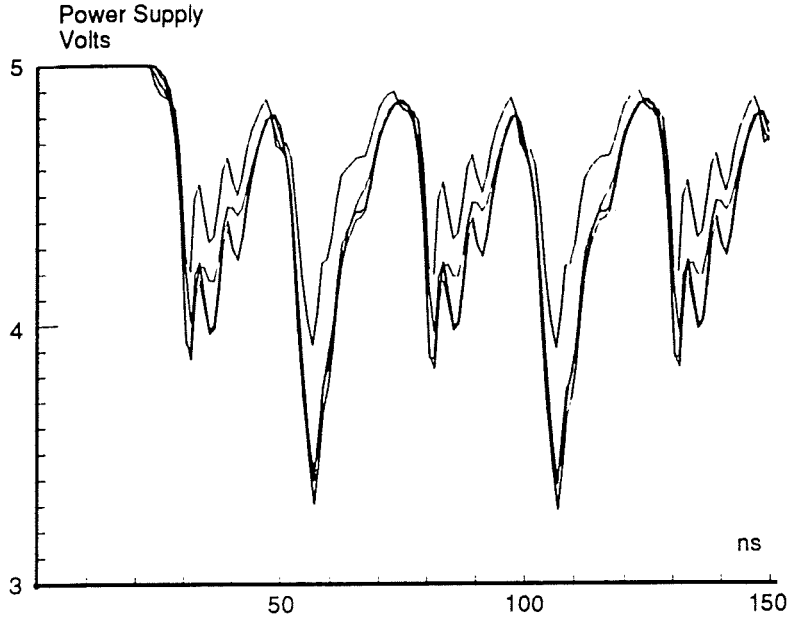
Storage limitations associated with the circuit simulator made it necessary to separate the resultant processor slice noise model into its constituent positive (Vdd) and negative (Vss) contributions.

4.6.2 Noise Predictions

Power distribution noise predictions, based on standard distribution technology, are shown in Figures 4.25, 4.26 and 4.27 for positions in the network which correspond to the multiplier unit and each of the arithmetic units. These results are for nominal operating frequency (20MHz) and with worst-case input vectors. Input vectors are offset as described in section 4.3.3.

It is clear that similar voltage integrity is predicted for the multiplier unit and each of the arithmetic units. Overall, though, peak voltages for the multiplier unit are about 0.2V worse than for the arithmetic units and, since this analysis is based on worst-case assumptions, the results for the multiplier unit will be taken as representative of voltage integrity levels anywhere within the "processor slice".

The precise extent to which the resultant voltage integrity affects circuit performance will be evaluated in section 4.6.4. It is worthwhile noting at this stage, however, the



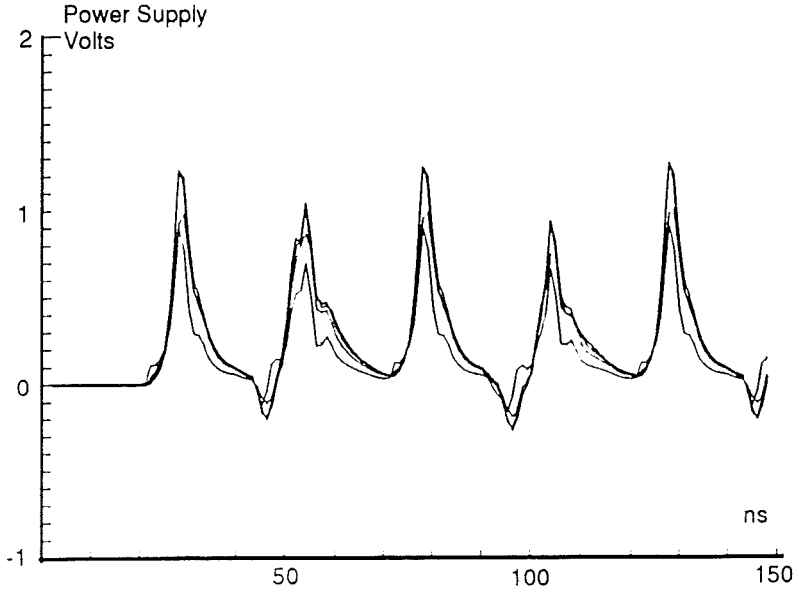
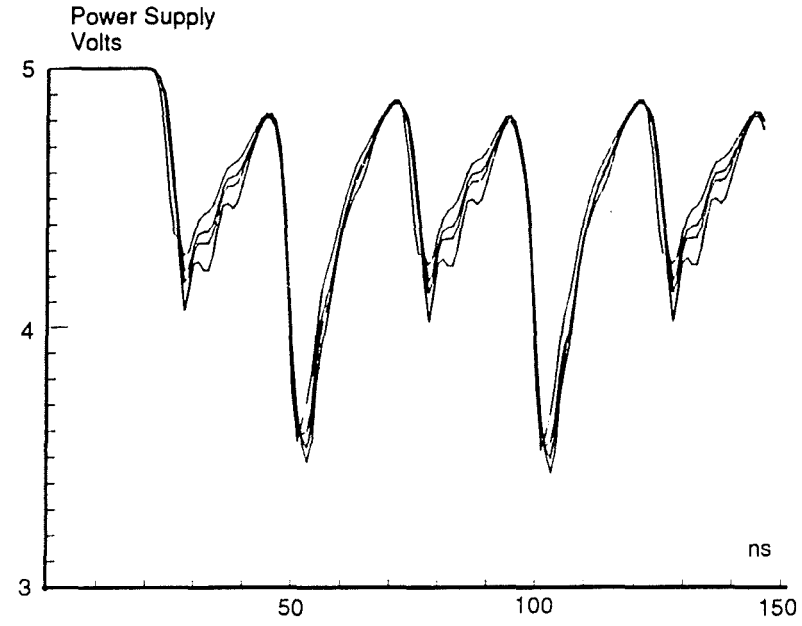


Fig 4.25 Noise Predictions for Multiplier Unit within "Processor Slice" (20MHz Operation)



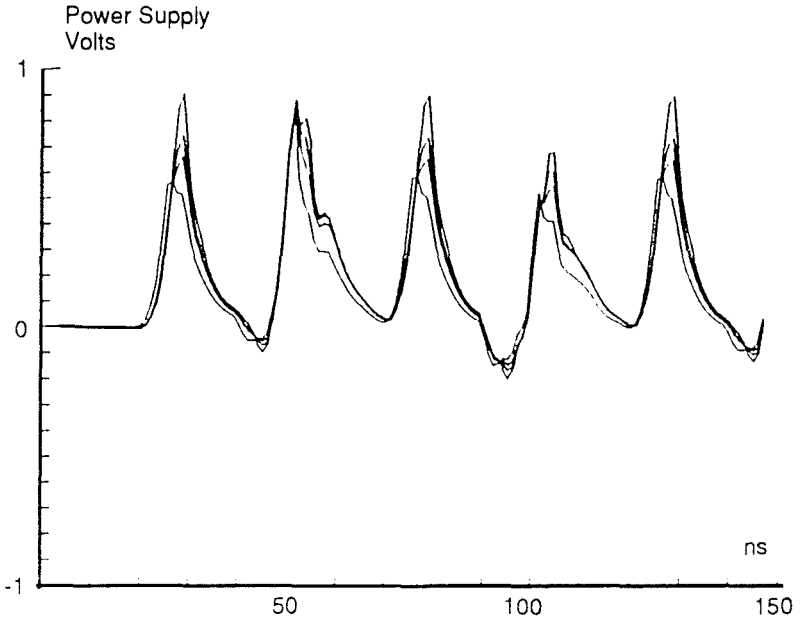


Fig 4.26 Noise Predictions for First Arithmetic Unit within 20 MHz Operation "Processor Slice

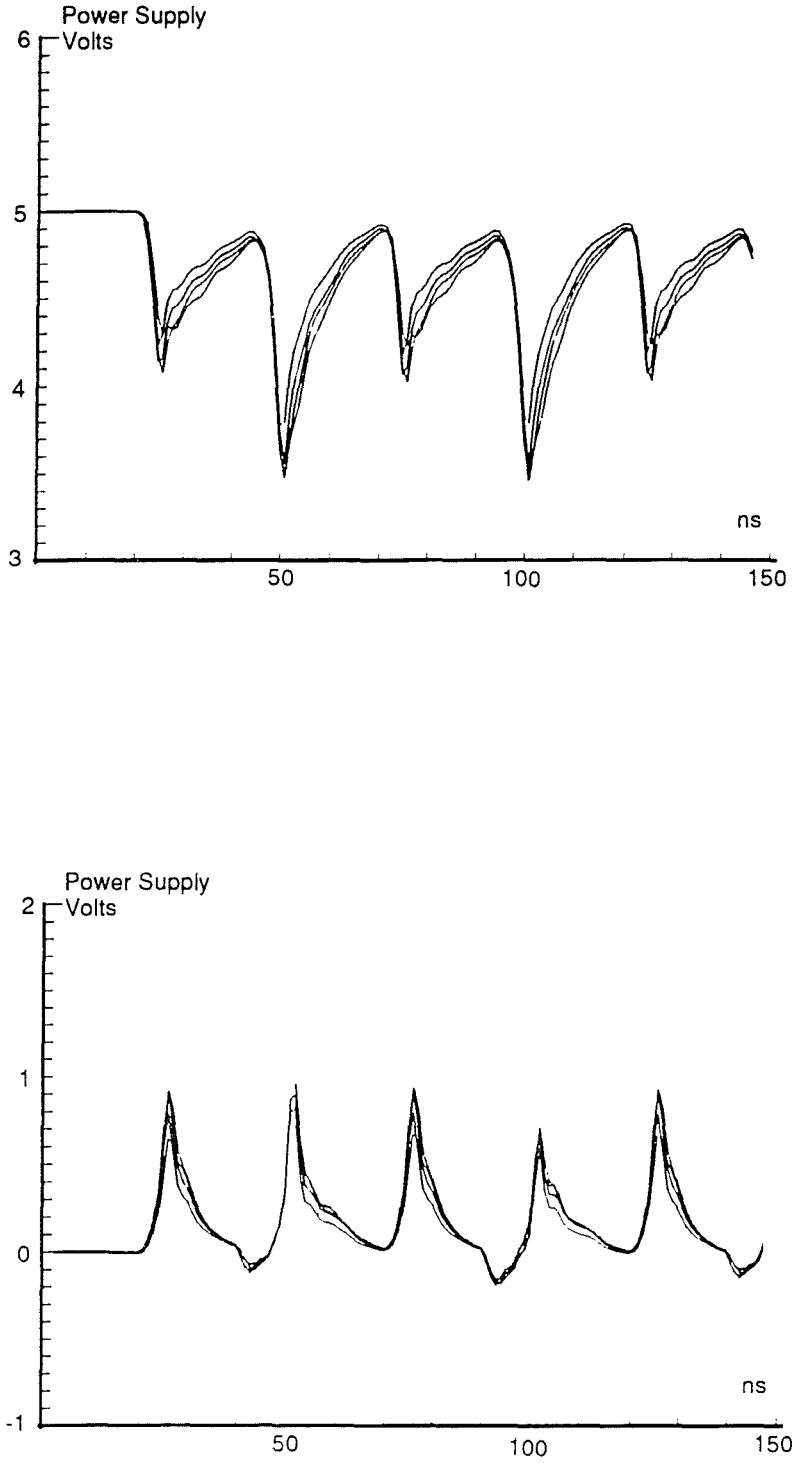


Fig 4.27 Noise Predictions for Second Arithmetic Unit within 20 MHz Operation "Processor Slice"

effect of not offsetting constituent circuit block activity by the latency of each sub-circuit. The noise predictions made for this case are shown in Figure 4.28 for positions in the network which correspond with the multiplier unit. If these are compared with the corresponding predictions shown in Figure 4.25, it is clear that peak values for each of the positive (Vdd) and negative (Vss) supplies are at least 0.3V worse if no offsets are applied. Since the corresponding change in voltage integrity, therefore, will be around 0.6V, the effect on circuit performance may be significant and will be evaluated in section 4.6.4.

In order to assess the effect of doubling the frequency of the "processor slice", the time between successive pulses in the voltage sources associated with each noise model was halved. The resultant predictions for the positive (Vdd) supply are shown in Figure 4.29 and without internal offsets in Figure 4.30. Their effect on circuit performance will be evaluated in section 4.6.4.

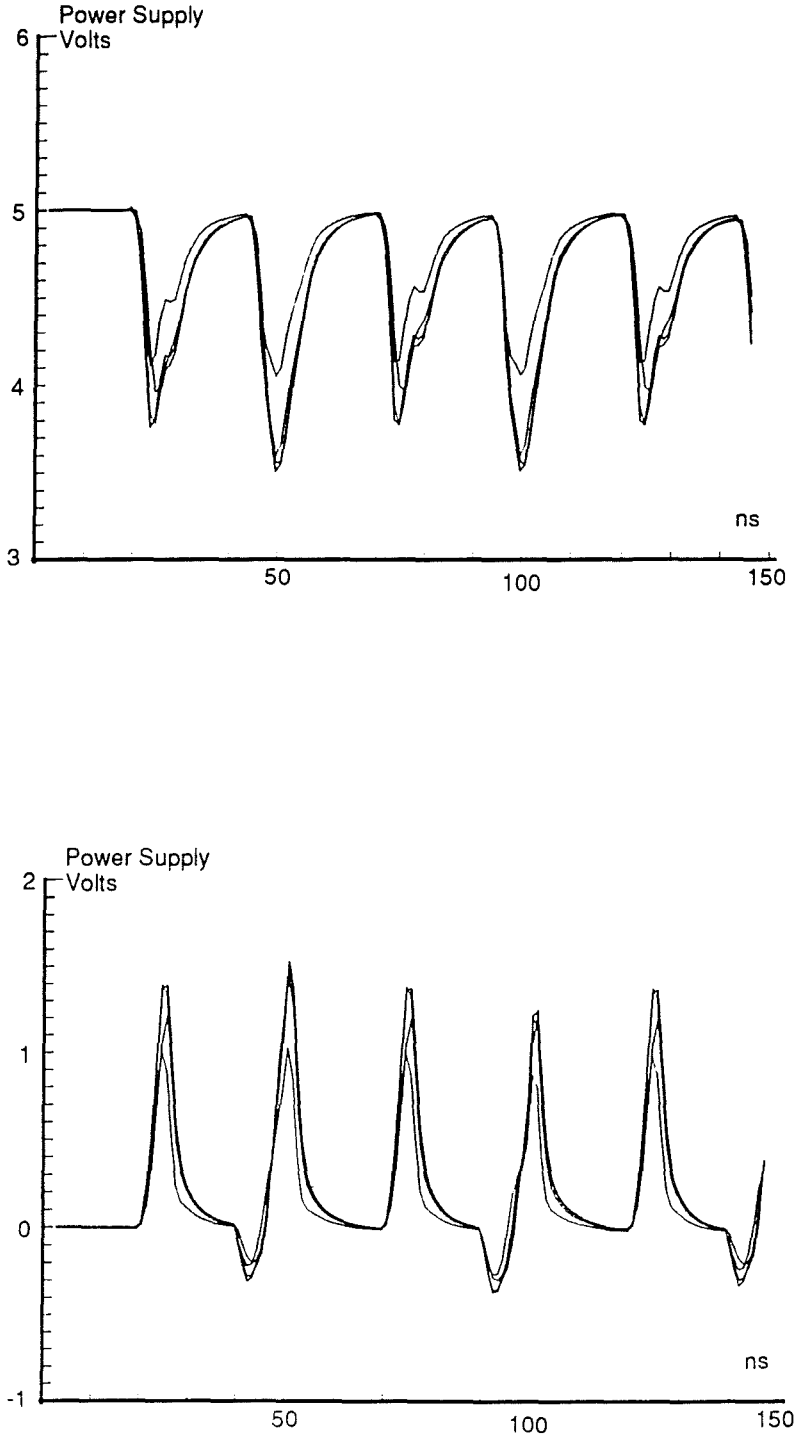


Fig 4.28 Noise Predictions for Multiplier Unit without Internal Offsets (20MHz Operation)

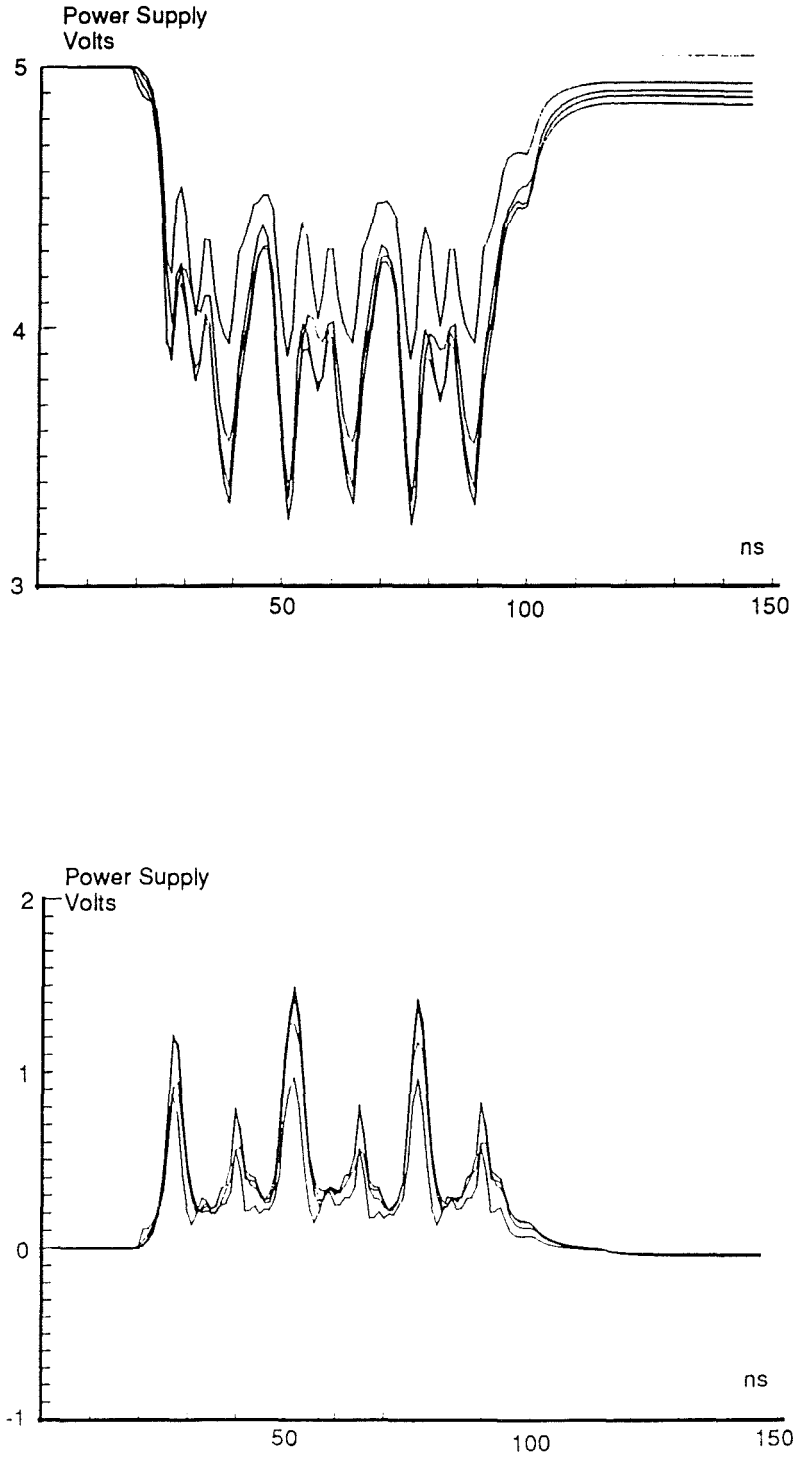
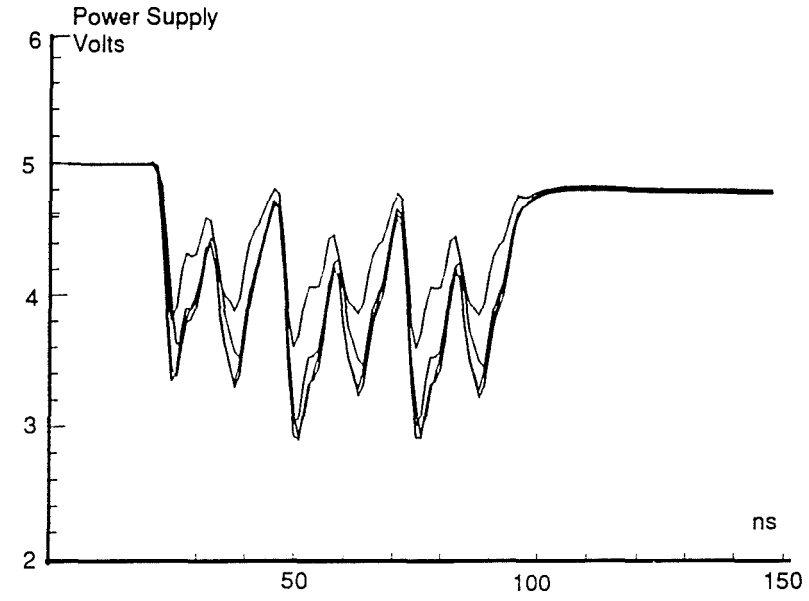


Fig 4.29 Noise Predictions for Multiplier Unit within "Processor Slice" (40MHz Operation)



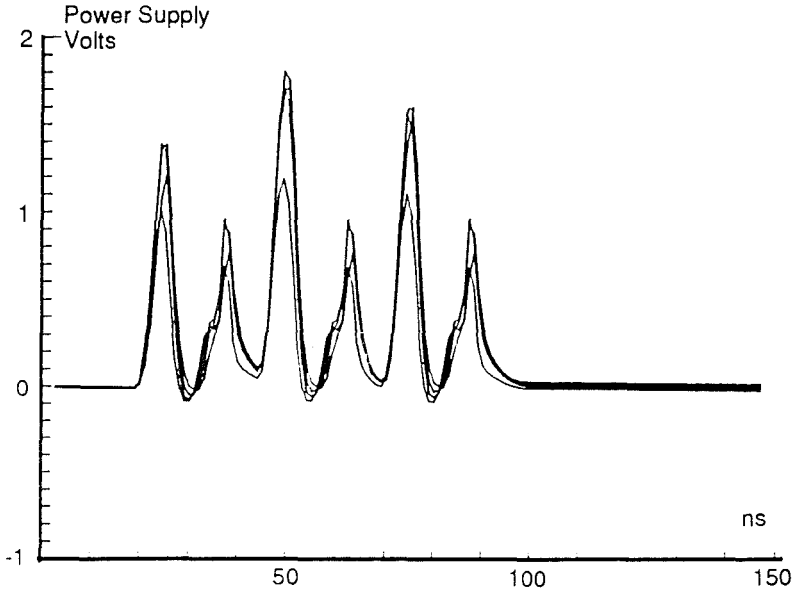


Fig 4.30 Noise Predictions for Multiplier Unit without Internal Offsets

4.6.3 Non-standard Technology

As in the analysis of chapter 3, predicted noise levels may be reduced if the electrical characteristics associated with non-standard, but emerging, power distribution technology were substituted for those of two-level aluminium metallisation with peripheral bond wire connections.

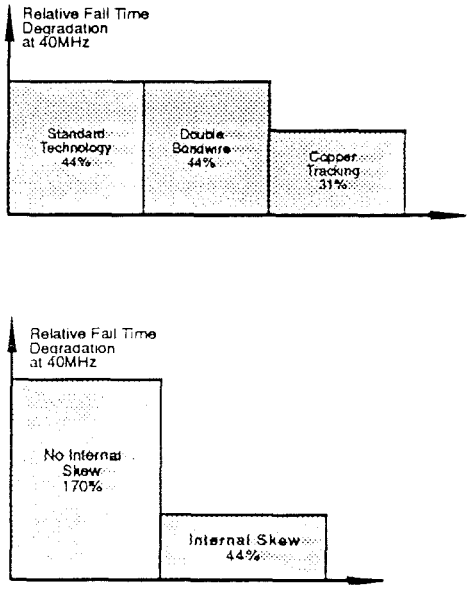
The non-standard technology variants, addressed in the systolic array analysis of chapter 3, are: double bond wire connections; electroplated copper tracking; and, non-peripheral bonding. With reference to Figure 4.24, it is clear that the distribution network for the "processor slice" renders inappropriate any analysis aimed at assessing the potential benefits of non-peripheral bonding.

Electrical parameters appropriate for double bond wire connections and for copper tracking of three microns thickness, were substituted in the simulation model for those associated with two-level aluminium metallisation with single bond wire connections. Their effect on circuit performance will be assessed in section 4.6.4.

4.6.4 Performance Implications

As is detailed in sections 3.3.0 and 3.3.1, the performance assessment methodology involves applying the predicted transient signal, for each of the positive (Vdd) and negative (Vss) supply lines, to the appropriate terminals of a CMOS inverter, thereby subjecting the inverter to the predicted level of voltage integrity. This was done for the "processor slice" operating at normal (20MHz) and double (40MHz) frequency and with each of the technology variants addressed above. The effect of internal offsets on circuit performance also was assessed.

The results obtained were converted into values for fall-time degradation and are shown as histograms in Figures 4.31 and 4.32.



Figs 4.31 & 4.32 Relative Fall Time Degradation for Processor Slice

Immediately it is clear, from these figures, that the level of fall-time degradation forecast for the "processor slice" operating at 20MHz is negligible, while, if constituent circuit block offsets are excluded from the analysis, this value is increased to 65%.

With operating frequency at 40MHz, fall-time degradation is increased to 45% with circuit activity offsets, and to 170% without.

It, therefore, can be concluded that: 1) since the simulation is for worst-case input vectors, the power distribution network developed for the processor slice is adequate; and 2), that constituent circuit activity offsets do influence significantly the predicted fall-time degradation values and therefore form an important part of the simulation model.

4.6.5 Noise Model Sensitivity

The purpose of this section is to assess the extent to which noise model predictions are dependent on the assumptions and design choices made during its development. Since there is no fundamental difference in the simulation method and resultant noise model the sensitivity analysis will complement the results which were concluded for section 3.5.1.

Model sensitivity to variations in: (1), first- and second-layer metal resistivity and capacitance; (2), first-layer metal resistivity and capacitance independently of second-layer; (3) package pin inductance; and (4), load as observed by the power distribution network. The results of this analysis are summarised in Figure 4.33. These are for independent variation of each parameter by -75% to +100%.

	V_{DD} (mV)	V_{SS}(mV)
Metal1 & 2 Resistivity	200	300
Metal1 & 2 Capacitance	200	200
Metal1 Resistivity	150	200
Metal1 Capacitance	70	70
Package Inductance	50	90

Fig 4.33 Sensitivity Analysis

As was found in chapter 3, it is evident that independent variation of first-layer metal resistivity reveals that two thirds of the voltage change is associated with this layer and one third with the second. The observed sensitivity, therefore, is explained by the fact that first layer metal resistivity is twice that of second layer.

The converse is true of metallisation capacitance.

Overall there are no unexpected dependencies inherent in the noise model.

4.7 Performance of Memory Reticle Power Distribution Network

4.7.1 Noise Model

In developing a power distribution noise model for the memory reticle, storage limitations as referred to in section 4.5 made it necessary to undertake the following simplifications.

As for the "processor slice", it was necessary to separate the noise models developed for the RAM, control, ROM and I/O ports, into their constituent positive (Vdd) and negative (Vss) components. In addition, it was necessary to partition the memory reticle as shown in Figure 4.34. The supply connections are representative of those that appear in the prototype device.

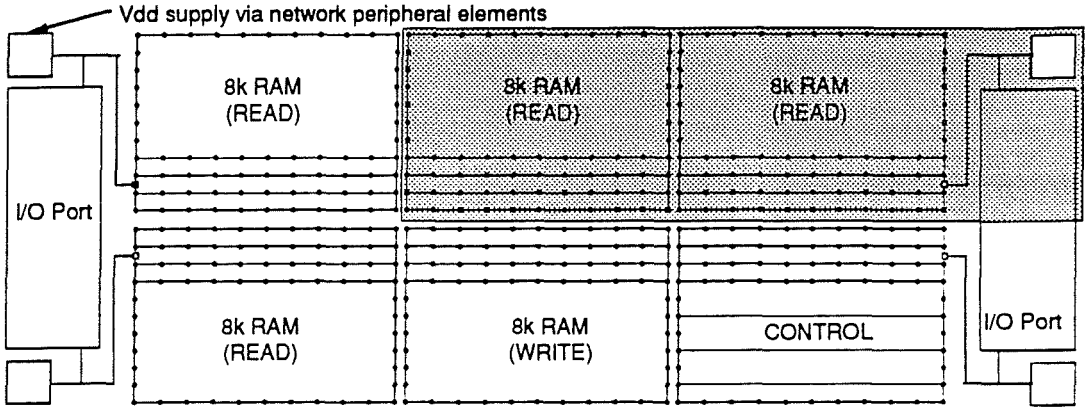


Fig 4.34 Memory Reticle Partitioning

The strategy behind the choice of partitioning is related directly to the various modes in which the device may operate. In order that the strategy be understood, these are described as follows.

The device has four modes of operation: (1), computing; (2), data-load; (3) data-unload data; and (4), "overlap" mode in which elements of the first three modes are present.

In mode (1), the computing mode, four of the eight 8k-RAM blocks associated with each memory reticle either are in "read" mode or in "write" mode. In mode (2), data-load mode, two of the eight RAM blocks are in "write" mode and the write port is active. In mode (3), data-unload mode, two of the eight RAM blocks are in "read" mode and the read port is active. In mode (4), "overlap" mode, two RAM blocks are in "read" mode, two are in "write" mode and both the read and the write ports are active. In modes (2), (3) and (4), the ROM blocks are inactive while in modes (2) and (3), the control block operates at 10MHz and not 20MHz.

It is clear from Figures 4.35 and 4.36 that, because of the RAM output drivers, the read cycle is overall more highly dissipating than the write. The I/O port is more highly dissipating when it is configured as an output or read port. Note also that the control block and ROM blocks dissipate much less than the RAM in either a read or

write cycle. The first mode of operation detailed above, therefore, will place greatest demands on the memory reticle power distribution network. This mode of operation will form the basis of this worst-case analysis and serves as a guide how best to partition the reticle into more manageable blocks.

With reference to Figure 4.2, it is clear that both the upper the lower memory reticles have four functional 8k-RAM blocks with one block used, in each case, for yield-enhancement Assuming that, due to processing defects, two 8k-RAM blocks in one reticle are non-functional and given that any of the ten 8k-RAM blocks may be accessed vias either I/O set, it is clear that greatest activity is effected when all five 8k-RAM blocks are operational with four out of five performing a read cycle and the fifth performing a write.

This level of activity was assumed.

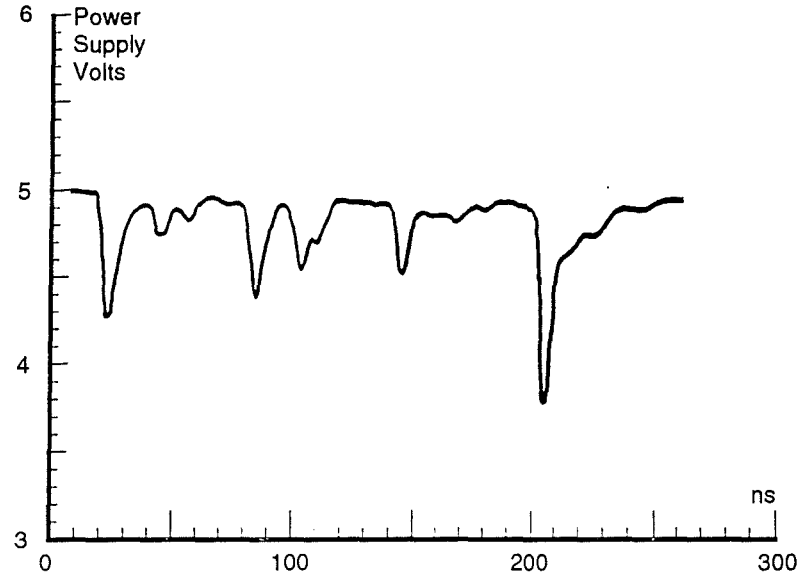
Bearing in mind each of the above facts and with reference to Figure 4.34, it is reasonable to assume that: (1), for reasons of symmetry about a horizontal line through the middle of the reticle, the worst-case circuit activity is obtained when the upper three RAM blocks are each performing a read cycle; and (2), for reasons of symmetry about a vertical line, that simulation of 2X8k-RAM blocks, as shown highlighted in the figure along with half an I/O port, will predict results for power distribution noise that are representative of the worst-case for the reticle.

Note that although the point of physical symmetry is represented by 1.5X8k-RAM blocks, the point of symmetry representative of circuit activity is offset to a value closer to 2X8k-RAM blocks because of the corresponding asymmetrical activity associated with the lower half of the reticle. Each of these simplifications was imposed by storage limitations associated with the circuit simulation software.

4.7.2 Noise Predictions

Noise predictions for the 2X8k-RAM and 0.5XI/O port during a read cycle and with standard power distribution technology are shown in Figure 4.37. These voltage integrity curves are for normal (10MHz) operating frequency and correspond to the positions, on the distribution network, shown in Figure 4.34.

Immediately, it is clear that, because of the more numerous supply connections, better voltage integrity is predicted for the negative network than for the positive.



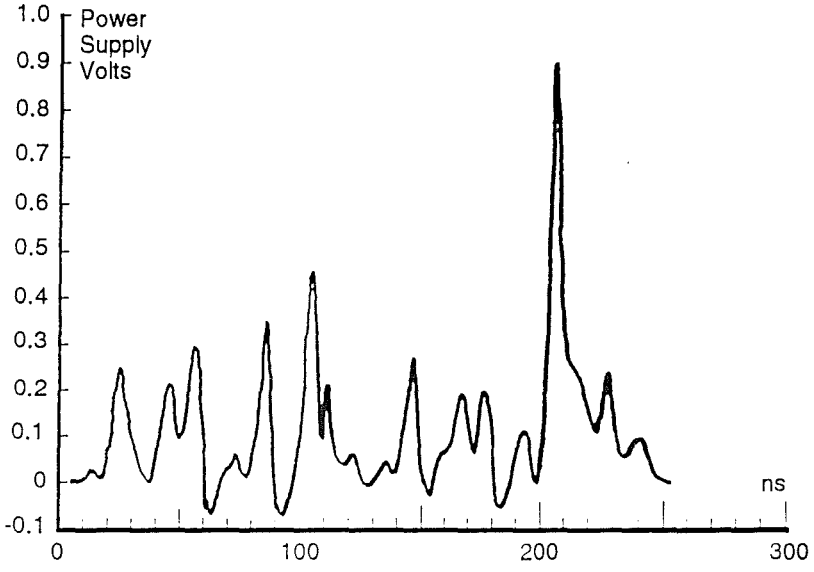
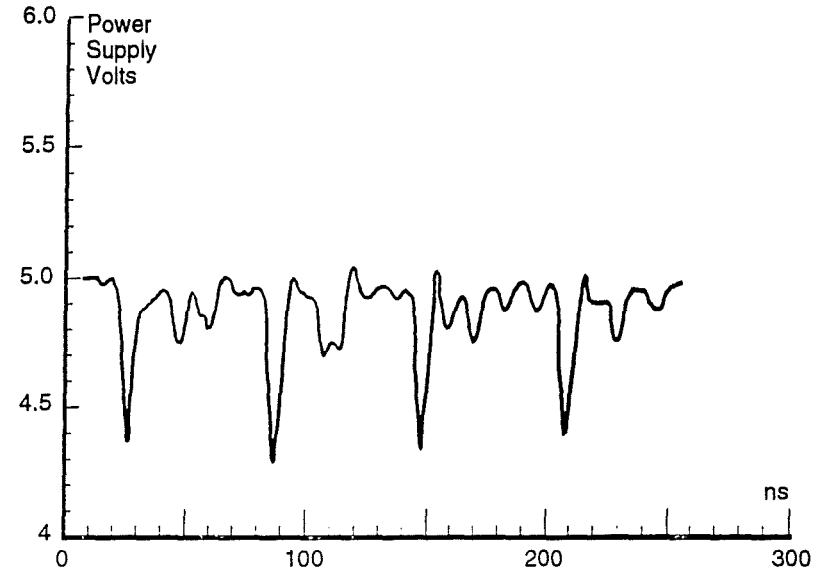


Fig 4.35 Noise Predictions for 8k RAM Block during Read Cycle



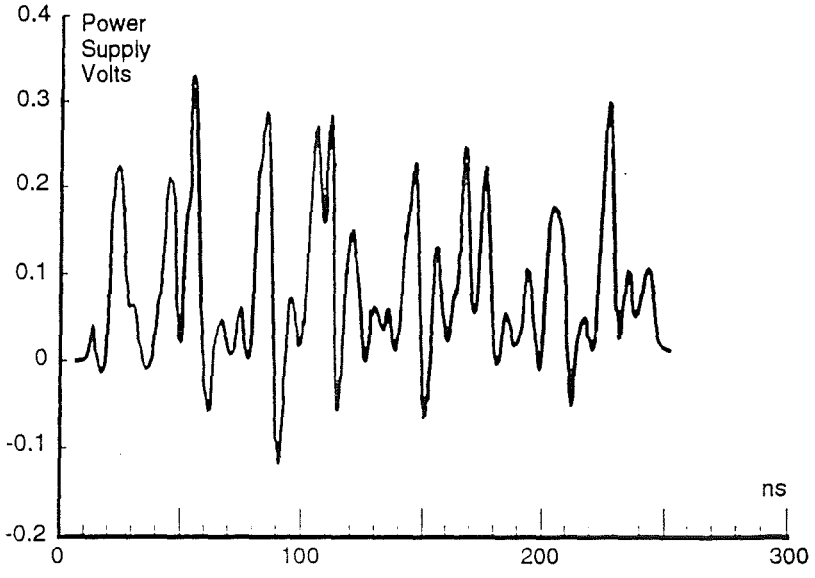


Fig 4.36 Noise Predictions for 8k RAM Block during Write Cycle

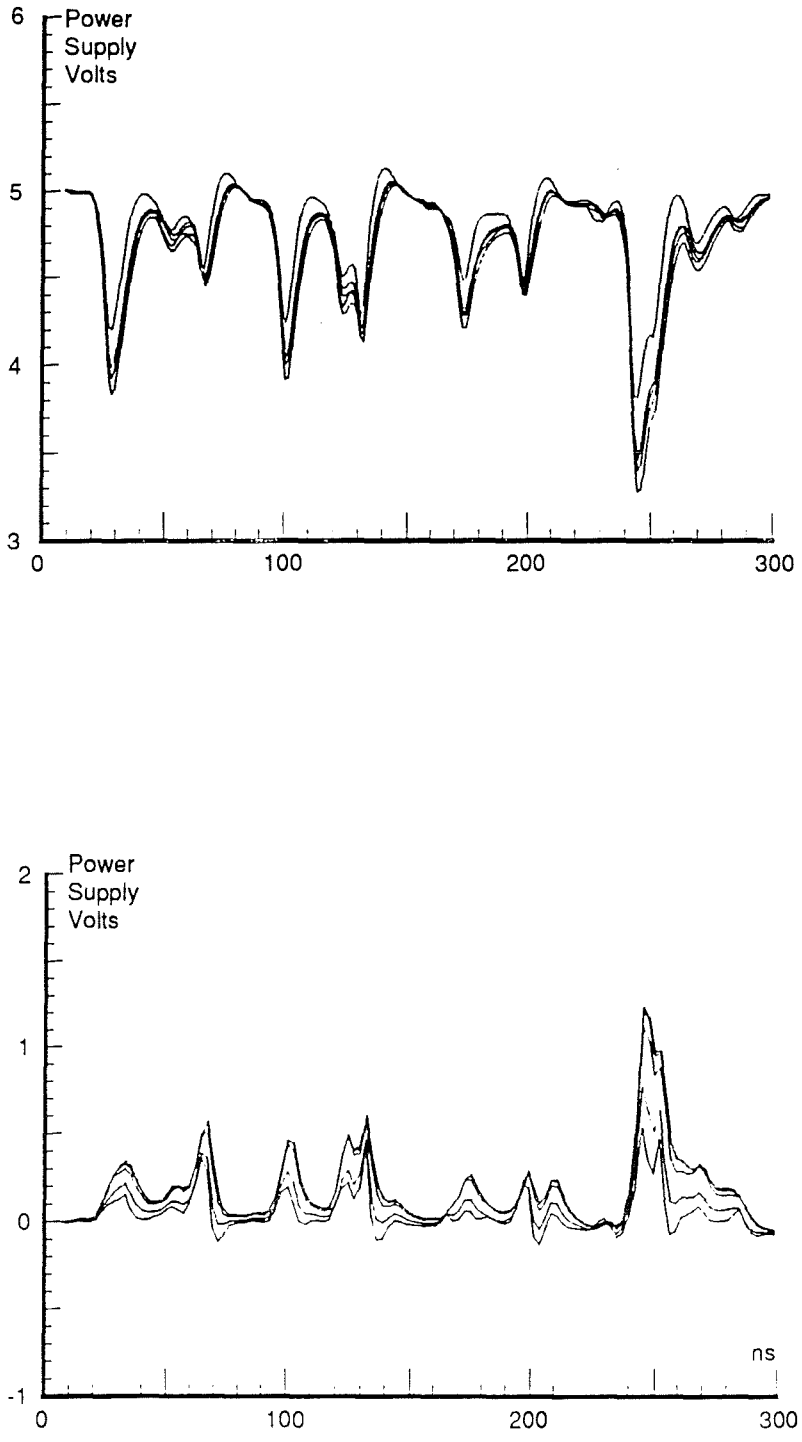


Fig 4.37 Noise Predictions for 8k RAM Block within Memory Reticle

4.7.3 Non-standard Technology

For similar reasons outlined in section 4.6.3, it was considered inappropriate to include non-peripheral bonding in this section. As in section 4.6.3, however, electrical parameters appropriate for double bond wire connections and for electroplated copper tracking were substituted for those of standard aluminium metallisation and single bond wire connections. Their effect on circuit performance will be assessed in section 4.7.4.

4.7.4 Performance Implications

The assessment methodology applied here is as detailed in section 4.6.4 and in sections 3.4.1 and 3.4.2 of chapter 3. It was applied for the RAM and I/O port combination operating at 10MHz and with each of the chosen power distribution technology variants. The results obtained were converted into values for fall-time degradation and are shown as histograms in Figure 4.38.

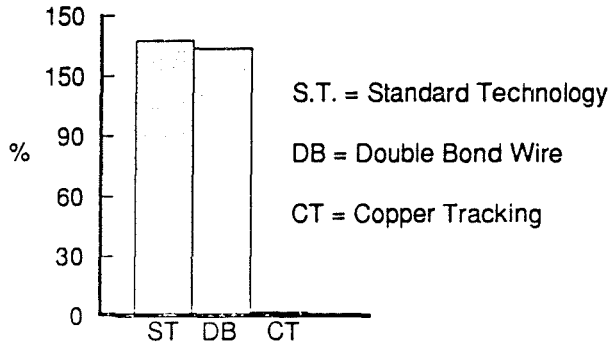


Fig 4.38 Relative Degradation for 8k RAM

The relatively complex structure of the predicted current flow waveforms, associated with the RAM and ROM circuit blocks rendered intractable the previously used method of generating current flow waveforms associated with double operating frequencies. For this reason, the analysis was undertaken at normal operating frequency only.

The figures suggest that, with standard operating frequencies and distribution technologies, fall-time degradation is 143% and while the improvements offered through double bond wire connections are negligible, those associated with electroplated copper tracking are such as to reduce fall-time degradation to zero.

4.7.5 Noise Model Sensitivity

This section complements section 4.6.5 in which a sensitivity analysis for the "processor slice" was undertaken. Model sensitivity to variations in: (1), first- and second-layer metal resistivity and capacitance; (2), independent variation of first-layer metal resistivity and capacitance; (3) package pin inductance; and (4), load as

observed by the power distribution network. The results are summarised in Figure 4.39.These are for independent variation of each parameter by -75% to +100%.

It is clear that the magnitude of the sensitivity for the memory reticle is approximately 50% higher than in the processor reticle. The nature of the sensitivity is the same.

4.8 Conclusions

4.8.1 Simulation Model

The simulation technique, developed in chapter 3, has been applied in the assessment

	V_{DD} (mV)	V_{SS}(mV)
Metal1 & 2 Resistivity	350	370
Metal1 & 2 Capacitance	300	300
Metal1 Resistivity	210	230
Metal1 Capacitance	120	120
Package Inductance	150	180

Fig 4.39 Sensitivity Analysis

of the extent to which power distribution noise may limit the performance of the above non-array-based signal processor circuit. The circuit requires over 600,000 active transistors and is designed to operate at 20MHz with 2.0 micron technology and 32MHz with 1.5 micron technology. Using 2.0 micron bulk CMOS technology, the circuit occupies 6.7 sq.cm.

This has been achieved by applying the techniques of chapter 3 to each circuit and then by composing a noise model for the sub-circuit blocks which, in turn, were combined to form power distribution noise models to represent each of the major components during their most active modes.

4.8.2 Performance Limitations

Relative gate-delay degradation for the "processor slice" is shown in Figure 4.31 and 4.32. As in chapter three, it is assumed that the introduction of logical errors will be avoided if a reduction in circuit performance is sustained apropos the various predicted increases in gate-delay. Armed with this assertion, the data can be used to assess the performance limit associated with each of the power distribution technology variants. In addition to the technology variants and operating frequencies, this was done for data derived with and without internal activity offsets. The results are summarised in Figure 4.40 and shown graphically in Figure 4.41.

It is clear that with 20MHz operating frequency, the "processor slice" is adequately supplied with current by the power distribution network. Only when the operating frequency is increased two-fold do the associated increases in gate-delay suggest a maximum operating frequency of 28MHz and around half that if internal activity

	Degradation
Standard Technology(20MHz)	None
Standard Technology(40MHz)	45%
3µm Copper Tracking(40MHz)	30%

Fig 4.40 Fall Time Degradation vs Technology

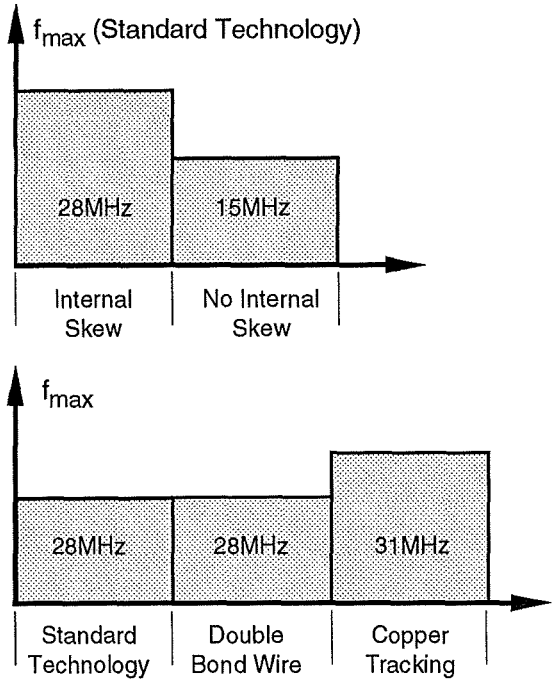


Fig 4.41 Maximum Frequency for Reliable Operation of "Processor Slice"

offsets are excluded. While there is no discernible improvement through the use of double bond wire connections, an improvement of around 15% is achievable with electroplated copper tracking.

Relative gate-delay degradation for the RAM and I/O port combination is shown in Figure 4.38. These data were used to assess an associated performance limitation in

	Degradation
Standard Technology(10MHz)	60%
3µm Copper Tracking(10MHz)	None

Fig 4.42 Fall Time Degradation vs Technology

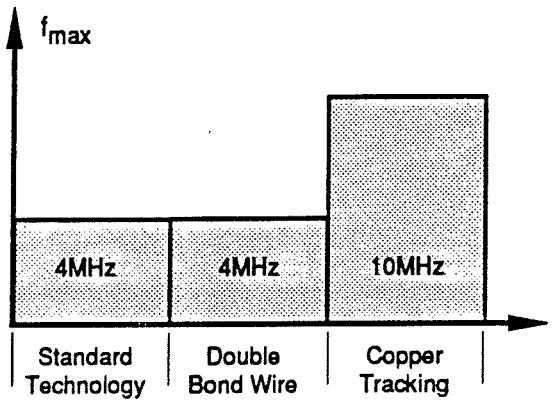


Fig 4.43 Maximum Frequency for Reliable Operation of 8k RAM Block

the now usual way. The results are summarised in Figure 4.42 and shown graphically in Figure 4.43.

It is concluded from Figure 4.43 that the power distribution network developed for the memory reticle has a current-carrying capacity which, for the RAM block, will incur a 60% reduction in performance.

As was the case for the "processor slice", it is clear that while no discernable improvements are predicted for double bond wire technology, an improvement of at least 60% is predicted for electroplated copper tracking. Note that in this latter case, the performance limit was not reached and hence the relative improvement is more.

This increased improvement over the processor reticle is explained by the higher sensitivity of the memory reticle distribution network to changes in resistance and capacitance and to the linear dependence of fall-time delay on voltage integrity.

The analysis has identified performance limits which, for the memory reticle, are 60% lower than the required design specification and, for the processor reticle, are 40% in excess of the required design specification. In addition, the analysis has predicted the extent to which non-standard technology variants can be expected to improve performance.

Other power distribution related issues which potentially may limit the size and performance of integrated circuits are those of latch-up and electromigration. In chapter 5, power distribution noise predictions are used to determine the relative susceptibility of the array-based and non-array-based architectures to the effects of transient latch-up and electromigration.

5 TRANSIENT LATCH-UP & ELECTROMIGRATION

5.1 Transient Latch-up

5.1.1 Introduction

Latch-up in bulk CMOS integrated circuits may be defined as a high current state accompanied by a collapsing or low supply voltage condition. Latch-up can be triggered by voltage or current transients at the power supply or output nodes, over-voltage conditions, or by photocurrents. It is a phenomenon which is now well understood with many references dedicated to its analysis [f01] and to techniques which render it much less likely [f02].

While most latch-up models [f03] are based on a strictly static analysis of the stability criteria, thereby ignoring its inherently transient nature, Troutman and Zappe [f02] propose a model which may be used to examine the current and voltage kinetics of a device during all phases which lead to the latched state.

Troutman and Zappe find that latch-up triggering depends not only on the parameter values of the model but also on the dynamics of the stimulus. The purpose of this analysis is to determine by simulation the extent to which the power distribution noise predicted in chapters three and four, effect circuit susceptibility to transient latch-up.

The transient latch-up model deliberately has been kept simple so as to focus on the effects associated with predicted power distribution noise. Emphasis is on qualitative understanding rather than quantitative accuracy.

5.1.2 Simulation Model and Methodology

This model is shown in Figure 5.1 with resistors R_w, R_s, R_{ep} and R_{en} representing the well-resistance, the substrate resistance, the p-epitaxial layer and the n-epitaxial layer. C_{sw} is used to represent the substrate-well capacitance and Q1 and Q2 to represent the lateral pnp and npn bipolar transistors. The transistors are represented by a basic Ebers-Moll model which incorporates an ideal exponential diode characteristic with reverse injected current flow.

It must be stressed that while all of the resistor values were chosen to be representative of modern CMOS processes (f05) and are as shown in Figure 5.1, the objective of this study is to determine comparative trends as opposed to absolute quantities. In short, their absolute value is unimportant.

Similarly, it is clear that in establishing trends rather than absolute quantities, the precise value of substrate-well capacitance and of alpha, the common base current gain, for each of the bipolar transistors is also unimportant. For the purposes of this analysis a value of 0.05 was assumed for each of the npn and pnp devices. This

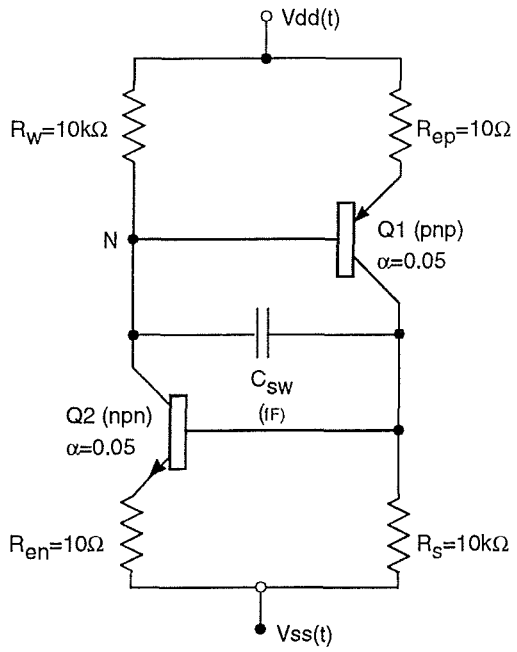


Fig 5.1 Transient Latch-up Model

independence of absolute values was exploited to determine the relative latch-up susceptability associated with integrated circuits of different size and performance.

By applying a piece-wise linear approximation of the positive (Vdd) and negative (Vss) supply lines to the appropriate terminals of the latch-up model, the substrate-well capacitance was increased until the potential at node N of Figure 5.1, the so-called "observed" supply voltage, collapsed thereby indicating the transient latched state. For the case of a single-array systolic array circuit operating at 20MHz, this effect is illustrated in Figure 5.2.

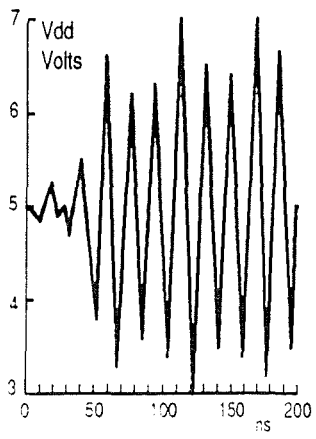


Fig 5.2(a)

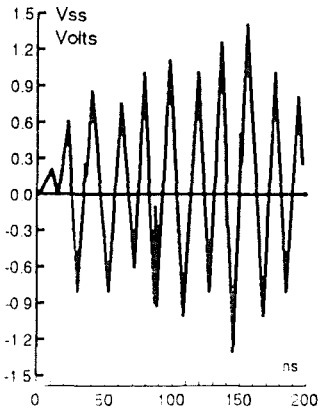


Fig 5.2(b)

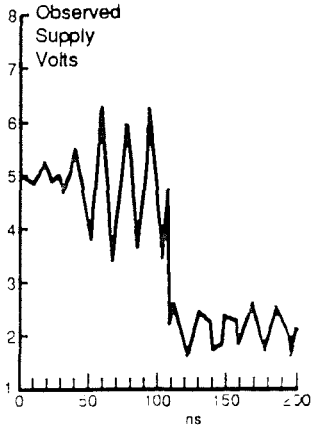


Fig 5.2(c)

Fig 5.2 Simulation of Collapsing Supply Voltage

In short, it is assumed that the value of substrate-well capacitance at which potential collapse occurs is a measure of the susceptibility of the circuit to transient latch-up.

5.1.3 Results

The exercise was undertaken for five-, eight-, thirteen- and seventeen-array systolic array circuits at operating frequencies of 10MHz, 20MHz and 30MHz.

The substrate-well capacitance values necessary to cause transient latch-up are shown in Figure 5.3. All other parameters associated with the latch-up model are as specified in Figure 5.1.

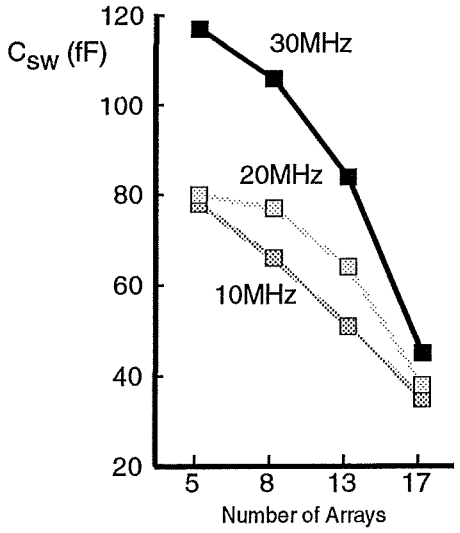


Fig 5.3 Substrate Well Capacitance necessary to cause Transient Latch-up

5.1.4 Conclusions

It is clear that transient latch-up susceptibility increases with larger circuits. As predicted in Chapters 3 and 4, the rate of change of supply voltage variation has a second-order linear form and that the second order nature is due only to packaging parasitics and not to metallisation inductance.

In addition, it was predicted that, for larger circuits, the second order exponential decay will gradually diminish due to increased source resistance. As the linear array is made larger therefore, the Ld^2i/dt^2 can be ignored and the i/C term will remain constant. Only R and di/dt will increase to cause the significantly higher dV/dt rates shown in Figure 5.4.

Consequently, as circuits become larger, more current will flow through the well (substrate) resistance (i=CdV/dt) and consequently induce a higher forward base-emitter bias on the lateral pnp (npn) bipolar device.

Conversely, it is predicted that, as circuit operating frequency is increased, then latch-up susceptibility is reduced. This is explained by the fact that at higher operating frequencies the voltage integrity associated with the terminals of the latch-up model is reduced so that the absolute voltage excursions associated with each positive (Vdd) and negative (vss) are reduced. This point is illustrated in Figure 5.5.

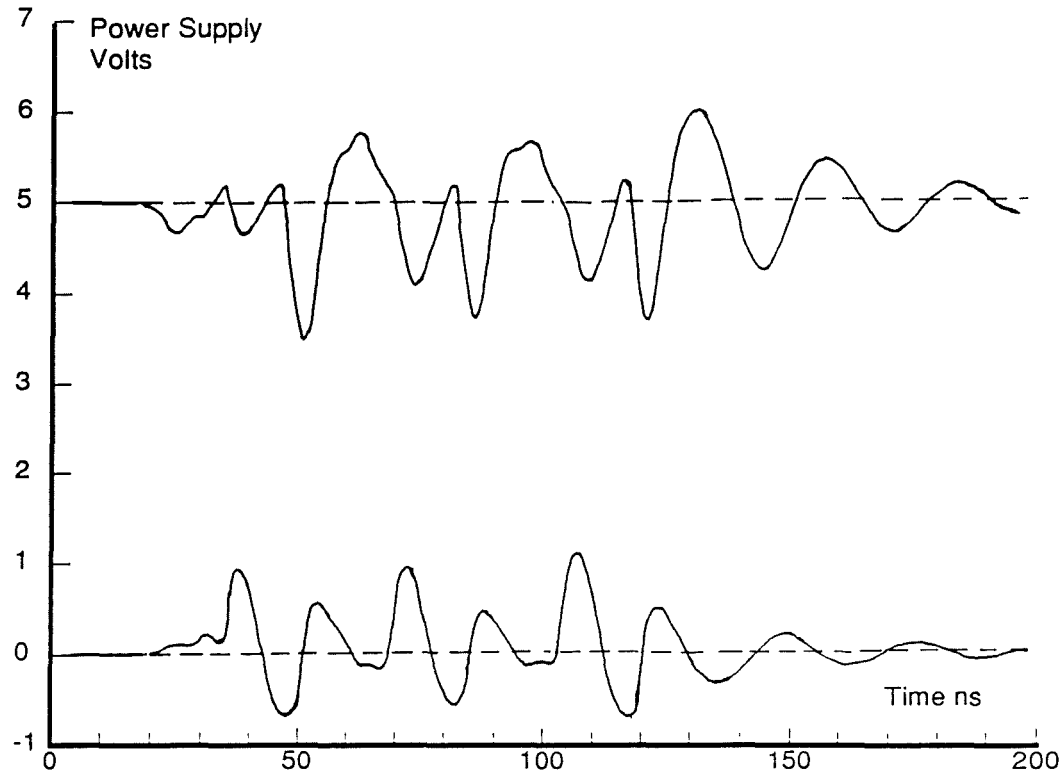


Fig 5.4(a) Five Arrays: 30MHz Operation

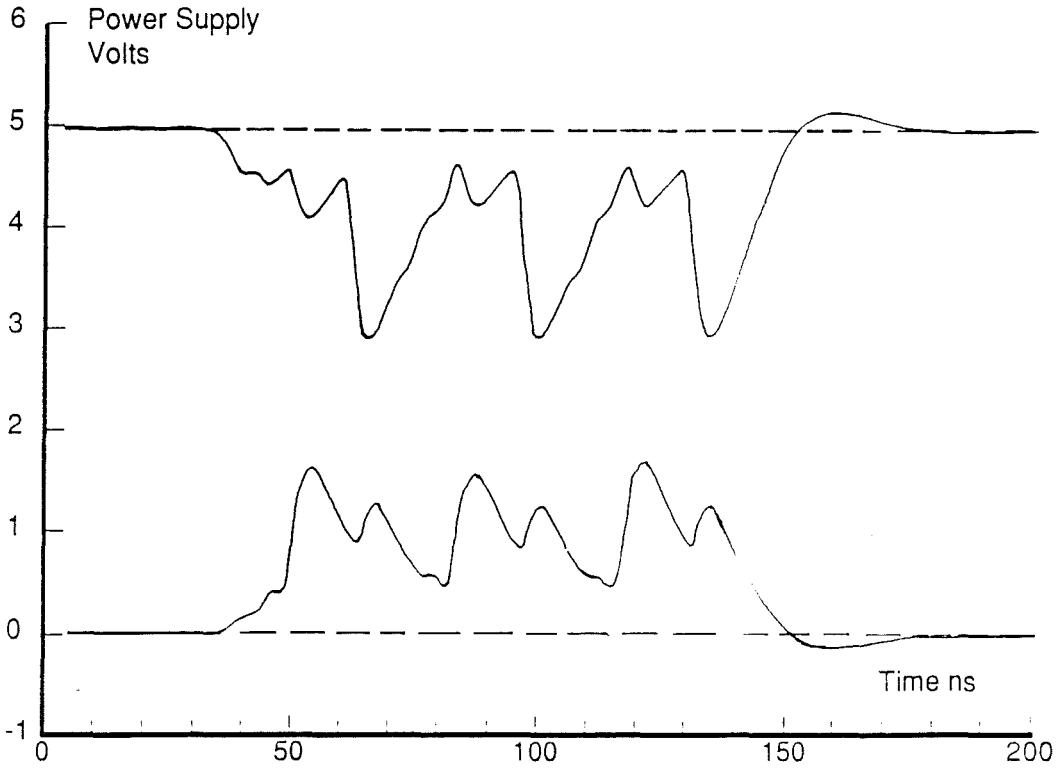


Fig 5.4(b) Seventeen Arrays: 30MHz Operation

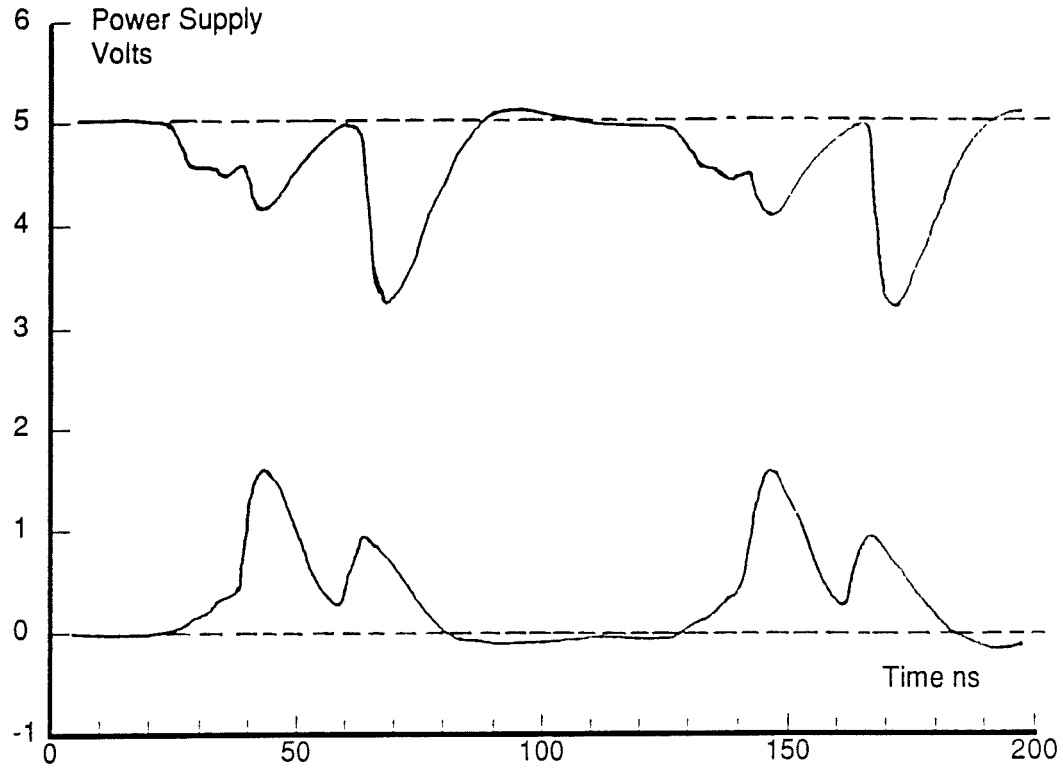


Fig 5.5(a) Seventeen Arrays: 10MHz Operation

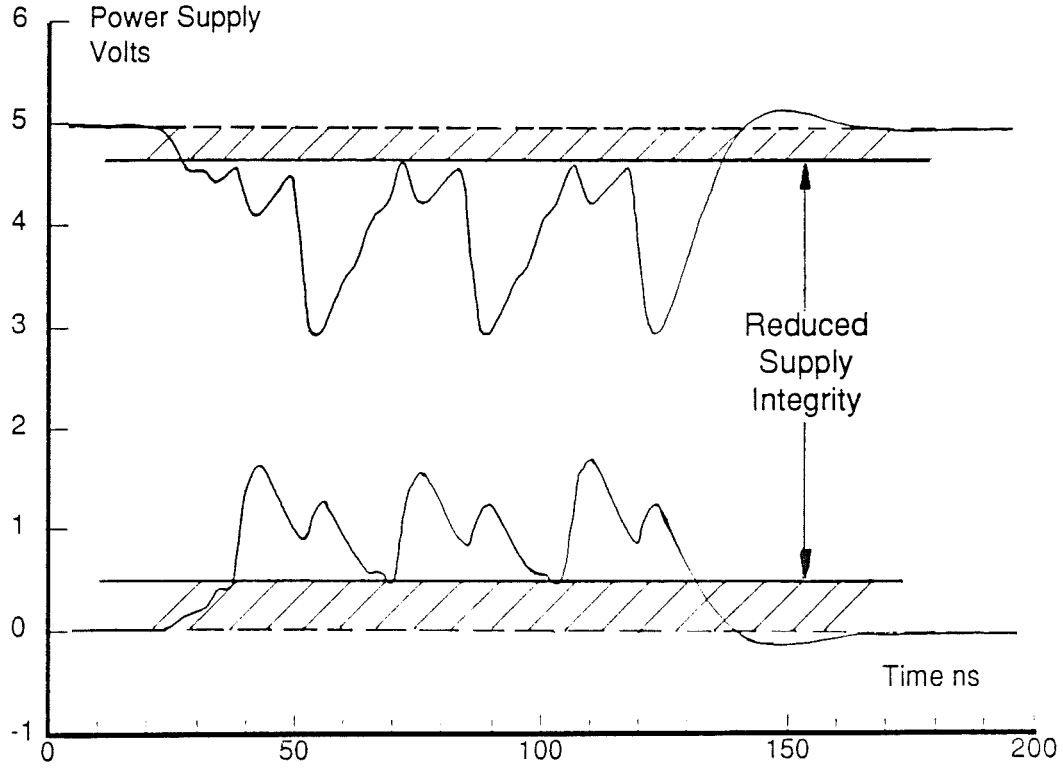


Fig 5.5(b) Seventeen Arrays: 30MHz Operation

5.2 Electromigration

5.2.1 Introduction

Electromigration is a term applied to the transport of mass in metals when stressed at high current densities. It occurs by the transfer of momentum from the electrons moving under the influence of the electric field associated with the conductor to the positive metal ions.

The electromigration resistance of aluminium metallisation can be increased by several techniques. These techniques include alloying with copper, the incorporation of discrete metal layers such as tungsten or encapsulating the conductor in a dielectric. The topic is well-researched with numerous references in the literature [f06], [f07], [f08].

The mean-time-to-failure (MTF) of the conductor can be related to the current density J in the conductor and an activation energy Q by,

$$MTF \propto Jx10^{-2}\exp[Q/kT] \qquad \text{(eqn 5.1)}$$

for $1.0X10^{-5}<J<2X10^{6}$ (A/sq.cm). The precise value for Q is dependent on grain size in the metal film, distribution of grain size and the degree to which the conductor exhibits fibre texture <111>.

Black [f06] describes two electromigration-related failure modes for integrated circuits utilizing aluminium metallisation. He concludes that for typical small conductor geometries, failures due to electromigration are such that the MTF will be reduced to less than ten years at current densities exceeding $5X10^{4}$ A/sq.cm and at temperatures in excess of 150°C.

5.2.2 Electromigration Analysis

It is assumed that this value of $5X10^{4}$ A/sq.cm is the threshold current density which, from a design-for-reliability viewpoint, ought not to be exceeded. This value will be used as a current density metric in the analysis that follows.

5.2.3 Results

Applying this metric to the distribution networks for the systolic array circuit, the processor slice and the memory block, results in the following maximum values for that metric.

1. Systolic Array	1.500A	(Metallisation layer-2 = 1.5mm wide)
2. Processor Slice	0.065A	(Metallisation layer-2 = 67 microns wide)
3. Memory/Read Port	0.030A	(Metallisation layer-2 = 32 microns wide)

With reference to Figure 5.6, it clear that the highest current for the seventeen-array systolic array circuit operating at 30MHz is around 0.8A and, therefore, is

significantly below the level required to cause an electromigration-related reliability hazard.

For the case of the processor slice operating at 20MHz, it is evident from Figure 5.7(a) that, operating at 20MHz, both the Vdd and Vss peak currents periodically exceed electromigration limits.

It is clear, from Figure 5.7(b), that the processor slice operating at 40MHz poses a reliability hazard. Further it is clear, from Figure 5.8, that the memory reticle power distribution network constitutes a considerable electromigration reliability hazard.

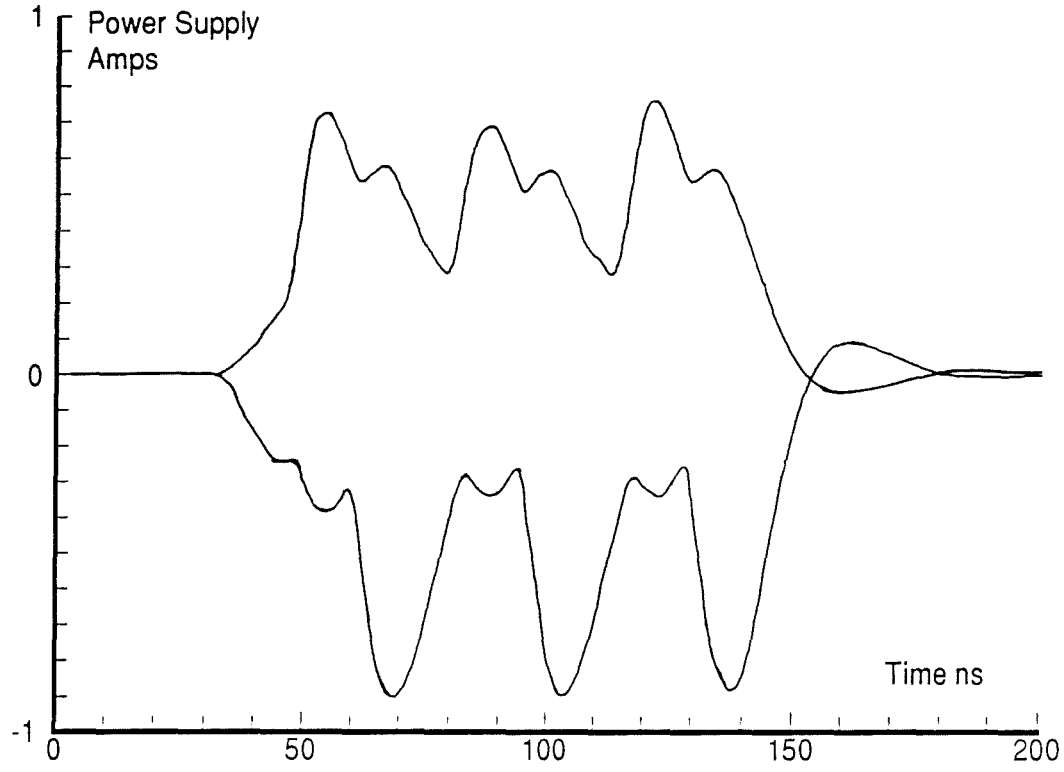


Fig 5.6 Seventeen Arrays: 30MHz Operation

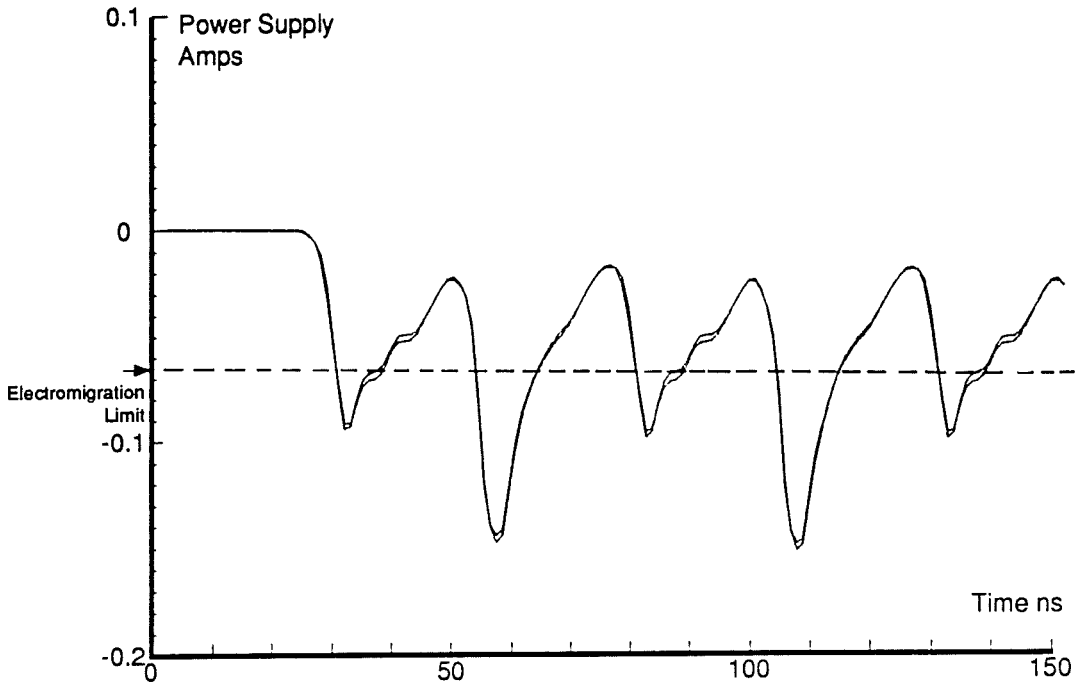


Fig 5.7(a) Vdd Supply Current for Processor Slice (20MHz Operation)

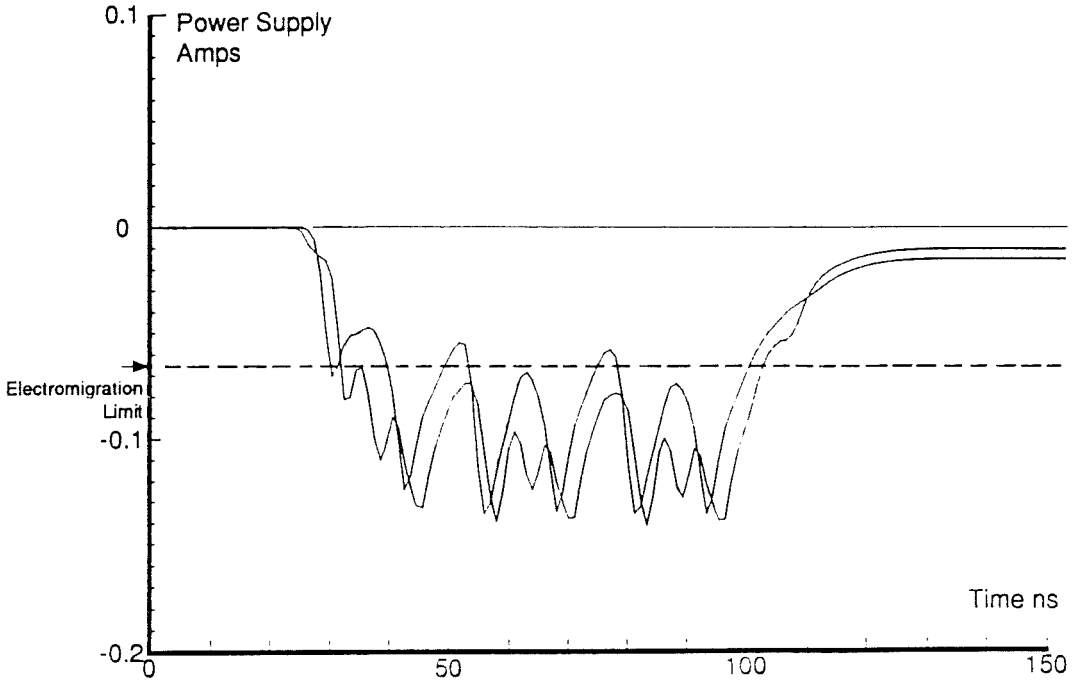


Fig 5.7(b) Vdd Supply Current for Processor Slice (40MHz Operation)

Aside from these specific findings, in which it would appear that power distribution electromigration becomes a reliability hazard after the onset of performance degradation, it is difficult to ascertain general trends regarding susceptibility of such highly synchronous circuits to electromigration.

Future work in this area could be aimed at determining the effect of prolonged and peak currents on the exponent values in Equation 5.1.

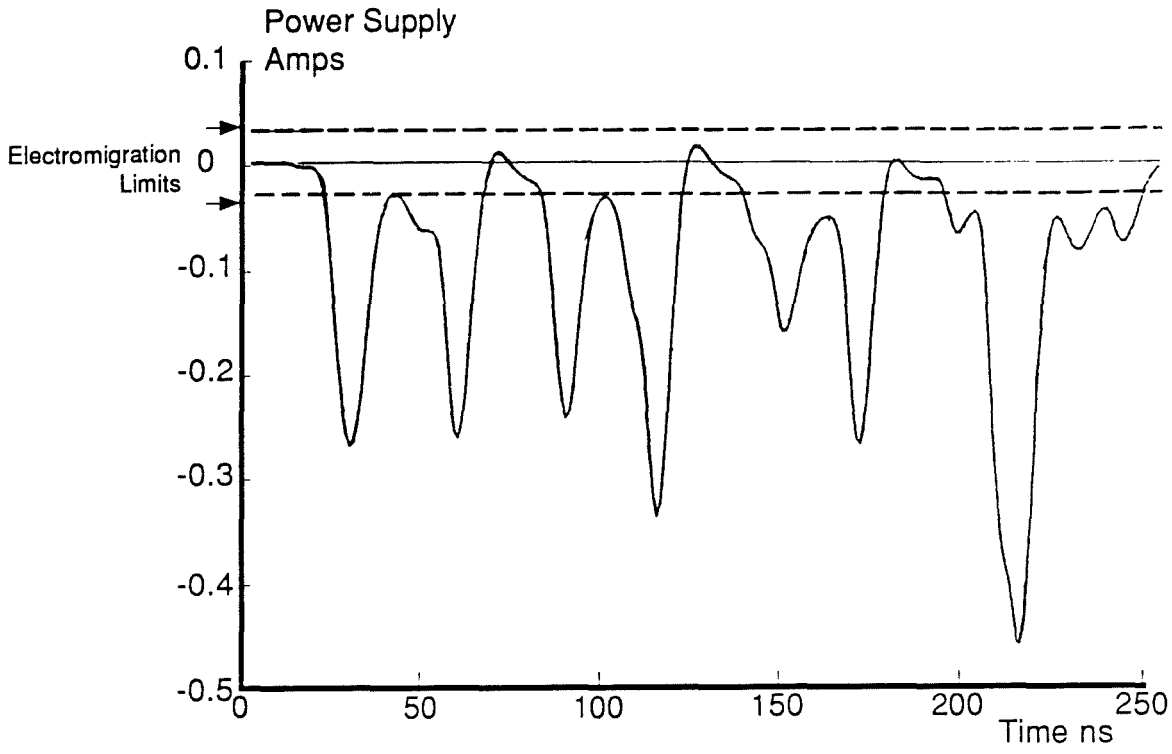


Fig 5.8 Vdd Supply Current for Memory Reticle

CONCLUSIONS

1. The Noise Modelling Methodology

A simulation methodology based on the SPICE circuit simulation programme has been developed to assess the magnitude, nature and effect of power distribution noise associated with systolic array integrated circuits of ever-increasing size.

The model adopts an equivalent circuit approach to noise modelling which has allowed a reduction in circuit simulation-related computational complexity by a factor of 15,000. It has been developed to assess the nature, the magnitude and effect of power distribution noise associated with systolic array integrated circuits which involve the integration of up to 1,020,000 transistors configured as a *linear* array of *LSI-sized* arrays.

The results may be extrapolated to the case where the array is physically square. These results then are representative of a systolic array occupying sixty-four sq.cm and involving some 14.5 million synchronously-active transistors.

The generated results provide new information regarding the magnitude, nature and effect of integrated circuit power distribution noise levels on circuit performance. The results are relevant to standard power distribution technology common to CMOS integrated circuits. In addition, the potential benefits associated with non-standard, but emerging, power distribution technologies have been assessed.

The simulation method has been developed for application to a non-array-based floating-point processor and RAM which contain around 600,000 transistors. The analysis has identified performance limits which, for the memory block, are 60% lower than the required design specification and, for the processor are 40% in excess of the required design specification.

A sensitivity analysis has revealed that the simulation model exhibits no instabilities consequent on the design choices and assumptions made during its development.

The effects of clock skew are predicted overall to be very small and insignificant at loads of below 165 Ohm-pF.

Transient latch-up susceptibility is predicted to increase with larger circuits. As circuits become larger, more current will flow through the well (substrate) resistance (i=CdV/dt) and consequently induce a higher forward base-emitter bias on the lateral pnp (npn) bipolar device inherent in the CMOS structure.

Conversely, it is predicted that, as circuit operating frequency is increased, then latch-up susceptibility is reduced. This is explained by the fact that at higher operating frequencies the voltage integrity associated with the terminals of the latch-up model is reduced.

2. The Nature of the Noise

The simulation model has revealed that the the distribution networks are second-order linear networks whose transient response is given by,

$$V(t) = \exp(st) = \exp(\alpha t).\exp(j\omega t), \quad \text{where } \alpha = -R_s/2Z_0$$

The attenuation factor, $\exp(\alpha t)$, results in decay periods which, for smaller LSI-sized circuits, are comparable with the clock periods chosen in this analysis. For larger circuits, the resistive nature of the network leads to attenuation factors ($\exp(\alpha t)$) which cause the associated sinewave component ($\exp(j\omega t)$) to decay significantly within the chosen clock period.

These facts result in a transient network response which, for smaller LSI-sized circuits, is dependent on circuit size and clock frequency. Such a response has been referred to as a *quasi-resonant* network response.

For larger circuits, where the attenuation factor ($-R_s/2Z_0$) is observed gradually to increase as distribution networks become larger, the power distribution noise has the form of a second-order differential equation in current with respect to time with L, C and R as equation coefficients.

$$dV/dt = Ld^2i/dt^2 + i/C + Rdi/dt$$

The form of the above equation has been borne out in a noise model sensitivity analysis which reveals no other instabilities.

The sensitivity analysis indicates no dependence on metallisation inductance thereby implying that the second order response of the distribution network is due only to package-related inductance and therefore can be capacitively decoupled externally. Internal recirculating currents, as described in section 1.5, are insufficient to cause any inductive supply noise.

Consequently, there is no long term problem associated with distribution network inductance. Package-related inductance may be decoupled externally [a03] and the lossy nature of metallisation interconnect prevents any second-order effects associated with metallisation inductance.

An indication of the potential benefits of copper-tracking, as described by Barrett [d11], is provided. Improvements in fall-time degradation of around 85% are predicted for electroplated copper tracking. Any investment in new low resistance, low inductance, interconnect technologies consequently should be welcomed.

The capacitative network decoupling techniques pioneered by IBM [c06], by providing more current internally, is an effective way of reducing external di/dt values and thereby reducing package-related noise.

3. A More Automated Approach

The equivalent circuit, of Figure 3.13, is the basic element around which the emergent simulation methodology is based. Future work should be directed at more automatically manipulating data associated with the initial circuit block simulations so that its form and syntax are compatible with that required for a SPICE piecewise linear voltage or current source.

After the initial simulation of individual circuit blocks, there is scope for a good deal of automatic data-formatting. The ultimate aim of this work would be a CAD tool which could be used more routinely to assess the performance of power distribution networks for highly synchronous digital CMOS integrated circuits.

4. Predicted Consequences

The third objective of this thesis was stated as an assessment of the extent to which power distribution noise limits the achievable level of synchronous circuit activity for CMOS digital integrated circuits. In the light of the results obtained for the devices involved in this study, it is likely that significantly higher data processing rates are achievable with digital CMOS integrated circuits only if new packaging and interconnect materials are developed thereby to sustain the power distribution requirements of increasingly synchronous circuits.

In a recent article [g01], M. Horowitz, Associate Professor at Stanford University, says of synchronous design styles ... "the big problem still remains the power needed to drive very fast clocks across large chips." Horowitz goes on to assert that another reason for avoiding synchronous design is that ... "clock and power distribution grids are especially difficult to analyse" and that dedicated clock and power interconnect layers fabricated with new low resistance materials probably will be needed.

In the same article, R. Krysiak, Design Manager for Inmos's T9000 Transputer, reveals that the T9000 clock distribution schemes had to be compromised because of the "difficulty of supplying the power required". Krysiak expects future Inmos designs to make use of locally synchronous clock distribution schemes. H.B. Bakoglu, Manager of IBM's Graphics VLSI Division warns that ... "design tools will have to become much more capable of extracting interconnect parameters".

Locally synchronous or partially asynchronous design styles though less demanding of the power distribution network result ultimately in compromised data-processing performance when compared with their fully synchronous counterparts. In short, such schemes are not long term solutions to achieving extremely high data processing rates.

Looking at the longer term, C. Mead [g02], Professor of Computer Technology at the California Institute of Technology, warns that a conventional digital signal processing approach, though accurate, is fast-approaching a technological limit and predicts that ... "it will not be the data processing ability which leads engineers to choose analogue, not digital, solutions but the problem of supplying power to digital chips consisting of

hundreds of millions of transistors". Mead goes on to suggest that ... "in the long term, the cost of computing systems will become directly proportional to the amount of power they consume".

Mead goes on to illustrate his point with the comparison of a modern microprocessor performing about 10^7 operations per second and consuming about one Watt of power or 10^{-7} Joules of energy per operation to the human brain which also consumes around one Watt but performs about 10^{16} operations per second. To reach that level of computation using silicon devices would require 10^7 Watts.

Mead suggest using the exponential transfer function of the MOSFET operating in the sub-threshold regime instead of its hitherto relatively crude application as a digital switch. This idea has formed the basis of Mead's recent work and has resulted in the development of an audio processing integrated circuit which mimics the human cochlea. Mead's device requires 30,000 transistors and requires only 10^{-11} Joules per operation an improvement of 10^4 over conventional digital signal processing techniques.

Ultimately, Mead proposes a Wafer Scale Integrated (WSI) signal processing device in which an entire silicon wafer is covered with these sub-threshold analogue devices thereby resulting in 10^{11} operations per second and consuming only around one Watt of power. Since such levels of power consumption are comparable with current CMOS digital integrated circuits, heat removal is not a problem leaving the issues of process yield-enhancement and circuit packaging as those which need to be addressed before Mead's concept becomes a practical reality.

REFERENCES

Introduction

[a01] W.R. Moore, A.P.H. McCabe and V. Bawa,
"Fault Tolerance in a Large Bit-Level Systolic Array",
Proc. First IFIP Workshop on Wafer Scale Integration, Southampton, pp.259-272, July 1985.

[a02] P. Ivey, M. Huch, T. Midwinter, P. Hurat and M. Glesner,
"Design of a Large SIMD Array in Wafer Scale Technology",
Proc. Second IFIP Workshop on Wafer Scale Integration, Egham, pp.75-85, September 1987.

[a03] R. Kenneth Keenan,
"Noise Aspects of Applying Advanced CMOS Semiconductors",
Harris Semiconductor (Harris) and The Keenan Corporation, March 1989.

[a04] R.M. Lea,
"Wafer Scale Integration: Motivation, Perspective and Potential",
Proc. Second IFIP Workshop on Wafer Scale Integration, Egham, pp.3-17, September 1987.

Chapter One

[b01] J.S. Kilby,
"Miniaturized Electronic Circuits",
U.S. Patent 3,138,743, June 1964.

[b02] T.R. Reid,
"The Texas Edison",
The Texas Monthly, pp.102-109 and 176-182, July 1982.

[b03] G. Moore,
"VLSI: Some Fundamental Challenges",
IEEE Spectrum, Vol.16, p.30, 1979.

[b04] G.H. Heilmeier,
"Microelectronics: End of the Beginning or Beginning of the End",
Proc. IEEE International Electron Devices Meeting, pp.2-5, December 1984.

[b05] J. Meindl,
"Theoretical, Practical and Analogical Limits in ULSI",
Proc. IEEE International Electron Devices Meeting, pp.8-13, December 1983.

[b06] L.A. Glasser and D.W. Dobberpuhl,
"The Design and Analysis of VLSI Circuits",
Addison-Wesley Publishing Company, Reading, Mass., 1985.

[b07] T. Sakurai and K. Tamaru,
"Simple Formulas for Two- and Three-Dimensional Capacitances",
IEEE Transactions on Electron Devices, Vol.ED-30, pp.183-185, February 1983.

[b08] A.E. Ruehli and P.A. Brennan,
"Accurate Metallization Capacitances for Integrated Circuits and Packages",
IEEE Journal of Solid State Circuits, Vol.SC-8, pp.289-290, August 1973.

[b09] A.E. Ruehli and P.A. Brennan,
"Capacitance Models for Integrated Circuit Metallization Wires",
IEEE Journal of Solid State Circuits, Vol.SC-10, pp.289-290, December 1975.

[b10] Robert J. Antinone and Gerald W. Brown,
"The Modelling of Resisitive Interconnects for Integrated Circuits",
IEEE Journal of Solid State Circuits, Vol.SC-18, No.2, pp.200-203, April 1983.

[b11] H. Hasegawa, M. Furukawa and H. Yanai,
"Properties of Microstrip Line on Si-SiO2 System",
IEEE Transactions on Microwave Theory and Techniques, Vol.MTT-19, pp.869-881, November 1971.

[b12] S. Seki and H. Hasegawa,
"Analysis of Interconnection Delay on Very High-Speed LSI/VLSI Chips Using an MIS Microstrip Line Model",
IEEE Transactions on Electron Devices, Vol.ED-31, pp.1954-1960, December 1984.

[b13] H.A. Wheeler
"Transmission-Line Properties of Parallel Strips Separated by a Dielectric Sheet",
IEEE Transactions on Microwave Theory and Techniques, pp.172-185, March 1965.

[b14] M. Caulton and H. Sobol,
"Microwave Integrated Circuit Technology - A Survey",
IEEE Journal of Solid State Circuits, Vol.SC-5, No.6, pp.292-303, December 1970.

[b15] I.T. Ho and S.K. Mullick,
"Analysis of Transmission Lines on Integrated Circuit Chips",
IEEE Journal of Solid State Circuits, Vol.SC-2, No.4, pp.201-208, December 1967.

[b16] K.C. Saraswat and F. Mohammadi,
"Effects of Scaling of Interconnections on the Time Delay of VLSI Circuits",
IEEE Journal of Solid State Circuits, Vol.SC-17, No.4, pp.275-280, April 1982.

[b17] H.B. Bakoglu and J.D. Meindl,
"Optimal Interconnection Circuits for VLSI",
IEEE Transactions on Electron Devices, Vol.ED-32, No.5, pp.903-909, May 1985.

[b18] C.L. Seitz,
"Self Timed VLSI Systems",
Caltech Conference on VLSI, pp.345-355, January 1979.

[b19] F.U. Rosenberger, C.E. Molnar, T.J. Chaney and T.P. Fang,
"Q-Modules: Internally Clocked Delay-Insensitive Modules",
IEEE Transactions on Computers, Vol.37, No.9, pp.1005-1019, September 1988.

[b20] C. Mead and L. Conway,
"Introduction to VLSI Systems",
Addison-Wesley Publishing Company, Reading, Mass., 1980.

[b21] T.J. Chaney and C.E. Molnar,
"Anomalous Behavior of Synchronizer and Arbiter Circuits",
IEEE Transactions of Computers, Vol.C-22, pp.421-422, April 1973.

[b22] H.J. Veendrick,
"The Behavior of Flip-Flops used as Synchronizers and Prediction of their Failure Rate",
IEEE Journal of Solid State Circuits, Vol.SC-15, pp.169-176, April 1980.

[b23] F. Anceau,
"A Synchronous Approach for Clocking VLSI Systems",
IEEE Journal of Solid State Circuits, Vol.SC-17, pp.51-56, February 1982.

[b24] R.M. Lea and J.N. Coleman,
"Clock Distribution Techniques for Wafer Scale Integration",
Proc. First IFIP Workshop on Wafer Scale Integration, Southampton, pp.46-53, July 1985.

[b25] H.B. Bakoglu, J.T. Walker and J.D. Meindl
"A Symmetric Clock-Distribution Tree and Optimized High-Speed Interconnections for Reduced Clock Skew in ULSI and WSI Circuits",
IEEE ISSCC 86 Digest, pp.118-122, October 1986.

[b26] "Bipolar Packaging Lags Technology",
Electronic Engineering Times, p.75, October 1988.

[b27] R. Bowlby,
"The DIP may take its Final Bows",
IEEE Spectrum, Vol.22, No.3, pp.37-42, June 1985.

[b28] C.L. Cohen,
"Japan's Packaging Goes World Class",
Electronics, pp.26-31, November 1985.

[b29] W. Andrews,
"High Density Gate Arrays Tax Utility, Packaging and Testing",
Computer Design, pp.43-47, August 1988.

[b30] A.H. Mones and R.K. Spielberger,
"Interconnecting and Packaging VLSI Chips",
Solid State Technology, pp.119-122, January 1984.

[b31] S. Oktay and H.C. Kammerer,
"A Conduction Cooled Module for High Performance LSI Devices",
IBM Journal of Research and Development, Vol.26, No.1, pp.55-66, January 1982.

[b32] R.C. Chu, U.P. Hwang and R.E. Simons,
"Conduction Cooling for an LSI Package: A One-Dimensional Approach",
IBM Journal of Research and Development, Vol.26, No.1, pp.45-54, January 1982.

Chapter 2

[c01] G. Ditlow and A. Brown,
"The Delta-I Simultaneous Switching Problem",
IEEE ISSCC 83 Digest, pp.337-339, October 1983.

[c02] Thaddeus Gabara and David Thompson,
"Ground Bounce Control in CMOS Integrated Circuits",
Proc. IEEE International Solid State Circuits Conference, pp.88-89, February 1988.

[c03] N.G. Ziesse, J.R. Werko, J.M. Dishman and W.O. Schlosser,
"Power Bus Transients in Very High Speed Logic Systems",
IEEE ISSCC 84 Digest, pp.110-115, October 1984.

[c04] M. Shoji,
"Electrical Design of BELLMAC-32A Microprocessor",
IEEE ISSCC 82 Digest, pp.112-115, October 1982.

[c05] Y. Itoh, K. Nakagawa, K. Sakui, F. Horiguchi and M. Ogura,
"Noise-Generation Analysis and Noise-Suppression Design Techniques in Megabit DRAM's",
IEEE Journal of Solid State Circuits, Vol.SC-22, No.4, pp.619-622, August 1987.

[c06] H. Schettler, W. Haug, K.J. Getzlaff, C.W. Starke and A. Bhattacharyya,
"A CMOS Mainframe Processor with 0.5-um Channel Length",
IEEE Journal of Solid State Circuits, Vol.SC-25, No.5, pp.1166-1177, October 1990.

Chapter 3

[d01] H.T. Kung,
"Why Systolic Architectures ?",
IEEE Computer, Vol.15, No.1, pp.37-46, January 1982.

[d02] J.V. McCanny and J.G. McWhirter,
"Bit-level Systolic Array Circuit for Matrix Vector Multiplication",
IEE Proceedings, Vol.130, Pt.G, No.4, pp.125-130, August 1983.

[d03] R. Christie,
"The MA7170 Systolic Correlator, Architecture and Applications",
Proc. First International Workshop on Systolic Arrays, Oxford, pp.113-122, July 1986.

[d04] F. Catthoor and H. deMan
"An Efficient Systolic Array for Distance Computation Required in a Video-Codec based on Motion-Detection",
Proc. First International Workshop on Systolic Arrays, Oxford, pp.141-150, July 1986.

[d05] W. Verbiest et al.,
"Specification for a Motion Detection Systolic Array Chip Set",
Internal Report BTMC, February 1985.

[d06] J.V. McCanny and J.G. McWhirter,
"Optimised Bit Level Systolic Array Circuit for Convolution",
IEE Proceedings, Vol.131, Pt.F, No.6, pp.632-637, October 1984.

[d07] L.W. Nagel,
"Spice 2: A Computer Program to Simulate Semiconductor Circuits",
Memo ERI-M520, University of California, Berkeley, California, May 1975.

[d08] C. Mead and L. Conway,
"Introduction to VLSI Systems",
Addison-Wesley Publishing Company, Reading, Mass., 1980.

[d09] C. Mead,
"Analog VLSI and Neural Systems",
Addison-Wesley Publishing Company, Reading, Mass., pp.183-184, 1989.

[d10] B.I. Bleaney and B. Bleaney,
"Electricity and Magnetism",
Oxford University Press, Oxford, pp.193-196, 1978.

[d11] J.J.Barrett
"Electroplated Copper Conductors in ESPRIT 824"
Proc. Second IFIP Workshop on Wafer Scale Integration,
Egham, pp. 227-239, September 1987

Chapter 4

[e01] A.K.J. Stewart,
"The Development of a Fault-Tolerant ULSI Signal Processor",
Proc. IEEE International Conference on Wafer Scale Integration, pp.245-255,
January 1989.

Chapter 5

[f01] A. Ochoa Jr., W. Dawes and D. Estreich,
"Latchup Control in Integrated Circuits",
IEEE Transactions in Nuclear Science, Vol.NS-26, pp.5065-5068, December
1979.

[f02] R.R. Troutman,
"Latchup in CMOS Technology: The Problem and its Cure",
Kluwer Academic Publishers, Boston, MA, 1986.

[f03] R.D. Rung and H. Momose,
"DC Holding and Dynamic Triggering Characteristics of Bulk CMOS Latchup",
IEEE Transactions on Electron Devices, Vol.ED-30, pp.1647-1655, December
1983.

[f04] R.R. Troutman and H.P. Zappe,
"A Transient Analysis of Latchup in Bulk CMOS",
IEEE Transactions on Electron Devices, Vol.ED-30, No.2, pp.170-179,
February 1983.

[f05] J.R. Black,
"Mass Transport of Aluminium by Momentum Exchange with Conducting
Electrons",
Proc. IEEE Annual Reliability Physics Symposium, IEEE Cat.7-15C58,
November 1967.

[f06] J.R. Black,
"Electromigration Failure Modes in Aluminium Metallization for Semiconductor Devices",
Proceedings of the IEEE, Vol.57, No.9, September 1969.

[f07] M.J. Attardo, A.H. Lanzberg, W.E. Reese and G.T. Wenning,
"Aluminium Electromigration in Long Stripes",
Proc. IEEE Annual Reliability Physics Symposium, December 1968.

Conclusions

[g01] S. Parry,
"Clock Skews Cause Concern in Fast IC's",
New Electronics, pp.15-16, June 1991.

[g02] R. Causey,
"Semiconductors 20 Years On",
Electronics Weekly, pp.16-17, May 1991.